Generating Non-English Synthetic Medical Data Sets

Lenart Dolinar
University College London
London, United Kingdom

Erik Calcina
Jožef Stefan International Postgraduate School
Jožef Stefan Institute
Ljubljana, Slovenia

Erik Novak
Jožef Stefan International Postgraduate School
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Using synthetic data sets to train medicine-focused machine learning models has been shown to enhance their performance, however, most research focuses on English texts. In this paper, we explore generating non-English synthetic medical texts. We propose a methodology for generating medical synthetic data, showcasing it by generating medical texts written in a non-English mixed language. We evaluate our approach with seven different language models and assess the quality of the data sets by training a classifier to distinguish between original and synthetic examples. We find that the `Llama-3` performs best for our task.

Keywords

Synthetic data, healthcare data, multilingual data, large language models, classification

1 Introduction

The healthcare domain produces a lot of medical data that can be used to train machine-learning models to help medical personnel. For example, a machine-learning model designed to perform Named Entity Recognition (NER) on electronic health records (EHRs) needs extensive labeled data sets to accurately identify medical terms like diseases, treatments, and patient details. However, the data contains a lot of personal information, and hospitals cannot share it freely due to data protection. In addition, there are not enough examples to train the models for some problems, such as those relating to rare diseases. Because of this, synthetic data is being used as a substitute to train the models.

Most synthetic data generation approaches focus on generating English texts. These usually utilize large language models trained on predominantly English documents retrieved from the web. However, there are few examples of using them to generate non-English texts. Furthermore, the language models have difficulties generating texts that do not reflect the distributions found in the training sample. This includes medical texts, which are usually not accessible to the general public.

This paper proposes a methodology for generating medical synthetic data using open-source large language models. We apply the methodology to a medical data set written in a non-English mixed language, where the Latin and non-Latin script is used interchangebly. We test it with seven large language models and assess performance by training a classifier to distinguish original examples from synthetic ones. Using the same prompt, we find that the open-source Llama-3 model best generates synthetic data that reflects the original data set.

The remainder of the paper is as follows: Section 2 presents the related work on generating synthetic data using large language models. Next, the proposed methodology is described in Section 3. The experiment setting is presented in Section 4, followed by the experiment results in Section 5. We discuss the results in Section 6 and conclude the paper in Section 7.

Information Society 2024, 10–14 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.4

2 Related Work

This section describes the related work, focusing on large language models and methods for generating synthetic data.

2.1 Large language models

Large Language Models (LLMs) are models that were trained to generate human-like texts based on an extensive process of training on vast amounts of data. Models, such as Llama 3 [2], GPT-4 [9], Aya 23 [3] and Mistral [7], are often easy to work with by providing an input textual prompt, based on which the models respond. The LLMs are helpful in specialized fields, such as medicine, since they can be fine-tuned on extensive data sets containing medical terms and concepts. This enables them to perform well in tasks such as medical synthetic data generation [12]. Despite that, they are sometimes unable to follow the instructions in the prompt accurately, leading them to hallucinate, i.e. confidently produce wrong responses [5].

In our experiments, we investigate the LLMs' performance in generating synthetic medical data given specific constraints and detailed prompts to simulate the original data set as best as possible.

2.2 Synthetic medical data generation

Recently, synthetic medical data, generated using LLMs, has been used to enhance the performance of models for solving different natural language processing tasks in medicine.

One work focuses on generating a synthetic data set of electronic health records of Alzheimer's Disease (AD) patients based on a label that is provided [8]. They find that the performance of their system for detecting AD-related signs and symptoms from EHRs improves vastly when trained on synthetic and original data sets as opposed to training the system only on the original one. Another work investigated using LLMs for extracting structured information from unstructured healthcare text [13]. By generating synthetic data using LLMs and fine-tuning the model, they significantly improved the models' performance for medical-named entity extraction and relation extraction tasks.

Most related works focus on English synthetic data due to scarce non-English training data and the dominance of English in medical terminology [6]. This paper focuses on generating non-English medical texts.

3 Methodology

This section outlines our research methodology. We first present the pre-processing of the data set, followed by describing the synthetic data generation process. Finally, we present the description of synthetic data set evaluation using a classifier. Figure 1 shows the diagram overviewing the proposed methodology.

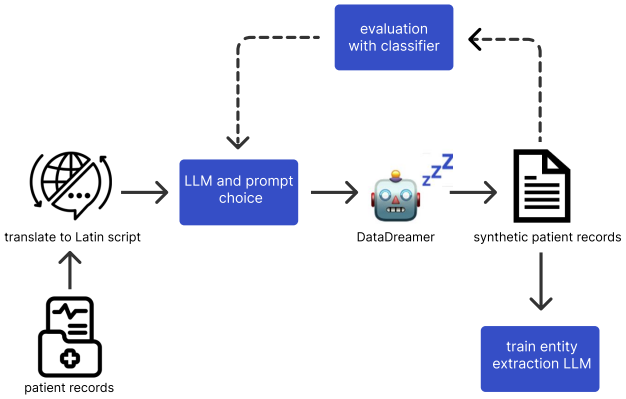

Figure 1: An overview of the methodology. The image was designed using resources from flaticon.

3.1 Data pre-processing

The data set used consisted of 1,299 examples of medical data written in a non-English mixed language, where the Latin and non-Latin scripts were used interchangeably. It also contained 1,495 labels, most of which were in English. The labels consisted of drugs, medical events, and measurements.

To translate the labels into the target non-English language, we used the `NLLB-200` [14] translation model[1]. Since LLMs were predominantly trained on texts written in Latin script, we decided to transliterate both the labels and examples from non-Latin to Latin script. This allowed the LLMs to generate longer tokens with richer information.

We split the original data set into two subsets to ensure no data leakage. The first one, consisting of 930 examples, was used for synthetic data generation. The second one, containing the remaining 369 examples, was used for evaluation.

3.2 Synthetic data generation

We utilized the datadreamer library [10] to generate the synthetic data set. The library enables open-source models to create synthetic data sets and was developed to work in research settings, supporting prompt templates and few-shot learning.

We developed a prompt containing the instructions and restrictions on generating the examples. To better showcase the structure of the generated text, we also provided five random examples from the original data set as few-shot examples. Next, using datadreamer, we sent the prompt to the chosen LLM. We experimented with multiple LLMs, and about 800 examples were generated for each used LLM. When experimenting with LLMs that required calling an external provider (e.g., OpenAI), we provided five static few-shot examples that did not include any patient personal data due to data privacy concerns.

To ensure the quality of generated data, we implemented a post-processing step. This included formatting the generated text into one line and excluding examples where the length was too long or where the model started repeating words meaninglessly. This ensured that all generated examples followed the same format and could be used for evaluation.

Analyzing the generated examples shows that they have many similarities. Therefore, there is a need for rigorous methods to evaluate how closely they resemble the original data set. The methods are explained in Section 4.1.

[1]https://huggingface.co/facebook/nllb-200-distilled-600M

3.3 Technical details

In this section, we describe the models and the parameters used in the experiment. All models used are available via the HuggingFace's transformer library [15].

We tested five open-source models to generate the synthetic data sets, all of which can be run on a 32GB GPU: `Llama-3` [2] only has support for the English language but has been fine-tuned to understand user prompts, which is a feature we expected would help a lot with the synthetic data generation.[2] `Aya-23` [3] is a multilingual language model and offers support for 23 languages, including the target non-English language.[3] `Mistral` [7] supports a variety of languages but omits the target language.[4] The models `Gemma-2` [4] and `Phi-3` [1] were also tested and compared in the experiments.[5,6] In addition, we experimented with `GPT-4o` [9] and `GPT-3.5-Turbo`, which are accessible via the OpenAI API.

All models were given the same prompt containing instructions that included (1) generating target non-English texts written in Latin script and (2) containing a label randomly selected from the original data set, (3) examples are supposed to be at most 6 words long, (4) should provide concise responses, (5) structured format (all text must be in a single line, must use // and commas as separators, and must be similar in format as the provided few-shot examples). To stress some more important instructions, some instructions were given in capital letters and were also repeated.

4 Experiment Setting

This section describes the experiment setting, which consists of the evaluation process and the metrics used to measure the approach's performance.

4.1 Evaluation approach

The quality of the generated synthetic data was measured in two parts. The first consisted of statistical measurements, such as calculating the average length of the generated examples and finding the proportion of examples that included the required labels. These statistics were then compared to the original data set.

The second part consisted of training a classifier to discern if the input text was from the original or from the synthetic data set. The data set used to train and evaluate the classifier involved 369 randomly selected synthetic examples and 369 examples from the original data set, transliterated into Latin script. We chose 5-fold validation as our classification procedure and calculated the mean performance across all trials.

The classifier was trained using the BERT [11] language model, specifically the `bert-base-multilingual-cased` variant[7]. The classifier was trained using the following parameters: batch size = 16, epochs = 3, and learning rate = 2e-5. The same parameters were used for all synthetic data sets.

4.2 Metrics

To assess the quality of the generated synthetic data sets, we used the F1 score as our main metric for evaluating the classifier's performance. The target value was 0.5; if the performance is greater than 0.5, the classifier can discern the original from the synthetic examples. Hence, the synthetic data does not reflect

[2]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[3]https://huggingface.co/CohereForAI/aya-23-8B
[4]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[5]https://huggingface.co/google/gemma-2-9b-it
[6]https://huggingface.co/microsoft/Phi-3-medium-4k-instruct
[7]https://huggingface.co/google-bert/bert-base-multilingual-cased

the original data set. If the performance is less than 0.5, the classifier has difficulties separating the synthetic from the original data, which can be because the synthetic data contains copies of the original examples. In addition to the F1 score, we measured the classifier's accuracy, precision, and recall, which are also reported.

5 Results

In this section, we present the results of our experiment. We first present the statistical results, followed by the classifier's evaluation.

5.1 Statistical analysis

Table 1 compares the synthetic data sets and the original one regarding label occurrence and average example length. The label occurrence is 1.000 in the original data set, as all examples from the original data set are assumed to include relevant labels and information.

The most aligned synthetic data set regarding label occurrence was generated using `GPT-4o`, followed by `Llama-3`. However, in terms of average example length, the data set generated using `Gemma-2` performed the best, followed by `Llama-3`.

The worst-performing models, in terms of label occurrence, were `Mistral` and `Phi-3`, which in about 25% did not include the selected label. The data set generated using the `Aya-23` had the largest difference in terms of average example length, on average generating examples with three extra words.

Table 1: Statistical comparison between the original and synthetic data sets. The bold and underlined values represent the best and second-best statistics, respectively.

LLM	Label occurrence	Avg example length
original data set	1.000	4.682
`Llama-3`	<u>0.990</u>	<u>5.330</u> (+0.648)
`Aya-23`	0.949	8.040 (+3.358)
`Mistral`	0.740	6.376 (+1.694)
`Gemma-2`	0.988	**4.207** (-0.475)
`Phi-3`	0.782	6.071 (+1.389)
`GPT-4o`	**0.996**	3.691 (-0.991)
`GPT-3.5-Turbo`	0.867	6.764 (+2.082)

Looking at both statistics, we can conclude that `Llama-3` had the best alignment to the original data set in terms of label occurrence and example length, closely followed by `GPT-4o`.

5.2 The classifier evaluation

Table 2 shows the F1, Precision, Recall, and Accuracy performances of the trained classifier on different synthetic data sets. The best performance was achieved by `Mistral` with approximately 0.85 scores in all four metrics, followed by `Llama-3`, with approximately 0.88 scores in all metrics. The worst performances were on data sets generated by the `Aya-23` and `GPT-3.5-Turbo` models. Surprisingly, the `Aya-23` is a language model supporting the target language; thus, it was expected to generate better examples.

6 Discussion

This section discusses the synthetic data generation performance, outlines our methodology's limitations and drawbacks, and proposes potential improvements to the approach.

6.1 LLM performance

Results in Table 1 show significant quality differences among synthetic data sets from different LLMs, with label occurrence ranging from 0.740 for `Mistral` to 0.996 for `GPT-4o`, and average example length from 3.691 for `GPT-4o` to 8.040 for `Aya-23`.

However, Table 2 indicates no significant performance differences within a single synthetic data set, with a maximal standard deviation of the metrics being 0.021 for the `Llama-3` data set.

We can also notice that the F1 and accuracy scores are very close for all synthetic data sets. This means the classifier was likely performing relatively similarly on both classes (synthetic and original data sets) without significant bias to either class.

We can observe much better performance on the `Llama-3` data set, which is primarily trained on English data, than on the `Aya-23` data set, which is also trained on the target language data. This shows that a model does not need to be extensively trained on non-English texts to generate this type of synthetic medical data well.

6.2 Limitations

Due to limited computing power, only one GPU with 32GB of space was available, restricting the testing of larger LLMs. To address these challenges, using cloud-based resources or distributed computing could help run larger models and improve the variety of synthetic data generated.

Due to privacy concerns, when using the OpenAI's `GPT-4o` and `GPT-3.5-Turbo` models, which are not locally-run models, we had to use five fixed examples when generating synthetic data instead of a larger variety. This potentially led to larger similarities of the `GPT-*` synthetic data sets to the examples instead of the original data set and, consequently, worse performance.

6.3 Potential improvements

The prompt was the same for all seven LLMs and was primarily tested on `Llama-3`. Hence, the performance might be biased towards the model. The method could be improved by tailoring the prompts to each model individually.

The evaluation of synthetic data sets could be further extended by checking for repeating examples in the synthetic data set or by checking how different the generated example is from the five provided examples. The evaluation could also be improved by checking for overfitting to the original data set.

7 Conclusion and Future Work

This paper presents a method for generating non-English synthetic medical data sets. To synthetically create data sets similar to the original, we carefully craft a prompt and perform preprocessing and post-processing of the data to increase performance and eliminate the effect of hallucinations.

Using a classifier and considering the inclusion of labels and generated text length, we conclude that `Llama-3` is best for generating examples that most closely resemble the original data set. In the future, we plan to explore the underlying architectures of the models to understand their performance differences in multilingual contexts. This will allow us to further refine our methods and create more accurate data sets.

Furthermore, we intend to use the synthetic data set to train a named entity recognition (NER) system to recognize medical labels from medical history examples. Measuring the performance of the NER trained on synthetic data sets will give us another way of evaluating their quality. We also intend to create a more general pipeline enabling the code to generate synthetic medical data in a wider variety of languages and formats.

Table 2: Mean performance metrics of the classifier for synthetic data sets, with standard deviation. Performances that are closer to 0.5 are considered better. The bold and underlined values represent the best and second-best performances, respectively.

LLM	F1	Precision	Recall	Accuracy
Llama-3	0.875 ± 0.021	0.881 ± 0.020	0.875 ± 0.020	0.875 ± 0.020
Aya-23	0.945 ± 0.005	0.947 ± 0.004	0.945 ± 0.005	0.945 ± 0.005
Mistral	**0.848 ± 0.012**	**0.856 ± 0.001**	**0.849 ± 0.011**	**0.849 ± 0.011**
Gemma-2	0.928 ± 0.005	0.930 ± 0.005	0.928 ± 0.005	0.928 ± 0.005
Phi-3	0.927 ± 0.009	0.932 ± 0.008	0.927 ± 0.009	0.927 ± 0.009
GPT-4o	0.906 ± 0.014	0.912 ± 0.012	0.907 ± 0.014	0.907 ± 0.014
GPT-3.5-Turbo	0.940 ± 0.013	0.944 ± 0.011	0.940 ± 0.013	0.940 ± 0.013

Acknowledgments

This work was supported by the Slovenian Research Agency. Funded by the European Union. UK participants in Horizon Europe Project PREPARE are supported by UKRI grant number 10086219 (Trilateral Research). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA) or UKRI. Neither the European Union nor the granting authority nor UKRI can be held responsible for them. Grant Agreement 101080288 PREPARE HORIZON-HLTH-2022-TOOL-12-01.

References

[1] Marah Abdin et al. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. 2024. arXiv: 2404.14219 [cs.CL]. URL: https://arxiv.org/abs/2404.14219.

[2] AI@Meta. "Llama 3 Model Card". In: (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[3] Viraat Aryabumi et al. *Aya 23: Open Weight Releases to Further Multilingual Progress*. 2024. arXiv: 2405.15032 [cs.CL].

[4] Google DeepMind Gemma Team. *Gemma 2: Improving Open Language Models at a Practical Size*. 2024. URL: https://storage.googleapis.com/deepmind-media/gemma/gemma-2-report.pdf.

[5] Xu Guo and Yiqiang Chen. *Generative AI for Synthetic Data Generation: Methods, Challenges and the Future*. 2024. arXiv: 2403.04190 [cs.LG]. URL: https://arxiv.org/abs/2403.04190.

[6] Rainer Hamel. "The dominance of English in the international scientific periodical literature and the future of language use in science". In: *AILA Review* 20 (Dec. 2007), pp. 53–71. DOI: 10.1075/aila.20.06ham.

[7] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[8] Rumeng Li, Xun Wang, and Hong Yu. "Two Directions for Clinical Data Generation with Large Language Models: Data-to-Label and Label-to-Data". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 7129–7143. DOI: 10.18653/v1/2023.findings-emnlp.474.

[9] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[10] Ajay Patel, Colin Raffel, and Chris Callison-Burch. *DataDreamer: A Tool for Synthetic Data Generation and Reproducible LLM Workflows*. 2024. arXiv: 2402.10379 [cs.CL]. URL: https://arxiv.org/abs/2402.10379.

[11] Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493.

[12] Karan Singhal et al. "Large language models encode clinical knowledge". In: *Nature* 620 (2023), pp. 172–180. DOI: 10.1038/s41586-023-06291-2.

[13] Ruixiang Tang et al. *Does Synthetic Data Generation of LLMs Help Clinical Text Mining?* 2023. arXiv: 2303.04360 [cs.CL]. URL: https://arxiv.org/abs/2303.04360.

[14] NLLB Team et al. *No Language Left Behind: Scaling Human-Centered Machine Translation*. 2022. arXiv: 2207.04672 [cs.CL]. URL: https://arxiv.org/abs/2207.04672.

[15] Thomas Wolf 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*. Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.