UTRad-NLP at #SMM4H 2024: Why LLM-Generated Texts Fail to Improve Text Classification Models

Yosuke Yamagishi
The University of Tokyo, Japan
yamagishi-yosuke0115@g.ecc.u-tokyo.ac.jp

Yuta Nakamura
The University of Tokyo, Japan
yutanakamura-tky@umin.ac.jp

Abstract

In this paper, we present our approach to addressing the binary classification tasks, Tasks 5 and 6, as part of the Social Media Mining for Health (SMM4H) text classification challenge. Both tasks involved working with imbalanced datasets that featured a scarcity of positive examples. To mitigate this imbalance, we employed a Large Language Model to generate synthetic texts with positive labels, aiming to augment the training data for our text classification models. Unfortunately, this method did not significantly improve model performance. Through clustering analysis using text embeddings, we discovered that the generated texts significantly lacked diversity compared to the raw data. This finding highlights the challenges of using synthetic text generation for enhancing model efficacy in real-world applications, specifically in the context of health-related social media data.

1 Introduction

In recent years, the burgeoning field of social media mining has opened new avenues for health-related research (Shakeri Hossein Abad et al., 2021), providing rich data sources for public health surveillance, including understanding public health trends and individual health behaviors. The Social Media Mining for Health (SMM4H) initiative, through its various tasks, aims to leverage these data sources to address pertinent health questions (Xu et al., 2024). This paper focuses on our approaches to Tasks 5 and 6, both of which present unique challenges and opportunities in the realm of text classification for health-related social media mining.

Task 5 targets the binary classification of tweets related to children's medical disorders, differentiating between tweets that report a genuine diagnosis and those that only mention these disorders without a diagnosis. Task 6 involves identifying the exact ages from social media posts, crucial for health research applications and enabling more accurate analysis of age-related health outcomes and behaviors in observational studies.

For both tasks, the challenge of imbalanced datasets is prominent. To address this, we employed the Large Language Model (LLM), GPT-4, aiming to augment our training data with synthetic positive examples to balance the dataset and enhance the performance of our binary classification models. Despite these efforts, our initial results were underwhelming, as the synthetic texts generated by GPT-4 lacked the diversity found in the raw data. This finding raises important questions about the practical challenges and limitations of using synthetic data augmentation in real-world applications, particularly in the nuanced field of health-related social media mining.

2 Dataset & Metrics

2.1 Task 5

The Task 5 dataset comprises tweets posted by users who reported their pregnancy on Twitter, used for binary classification. It includes 7,398 tweets for training, 389 tweets for validation, and 1,947 tweets for testing. The evaluation is based on the F1-score for tweets reporting a child with a disorder (annotated as '1').

2.2 Task 6

Task 6 focuses on extracting self-reported exact ages from posts on Twitter and Reddit. The dataset features 8,800 labeled tweets and 100,000 unlabeled Reddit posts from r/AskDocs containing 2-digit numbers for training; 2,200 tweets and 1,000 Reddit posts on dry eye disease for validation; and 2,200 tweets, 2,000 Reddit posts on dry eye disease, and 12,482 posts on social anxiety with ages 13 to 25 for testing. The evaluation uses the F1-score on the positive class ('1'), with micro-averaging.

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 42–47

2.3 Label Distribution

Both Tasks 5 and 6 have imbalanced datasets where Label 1 is less than half as numerous as Label 0. The distributions are documented in Table 1.

Table 1: Label Distribution in Tasks 5 and 6

Task	Label 0	Label 1
5	5,118 (69.1%)	2,280 (30.8%)
6	5,966 (67.8%)	2,834 (32.2%)

3 Methods

3.1 LLM Text Generation

We used the OpenAI API's GPT-4 ("gpt-4-0125-preview") with in-context learning techniques. We explained the label definitions and showed 5 randomly selected examples for each of the positive and negative labels. The model's temperature was initially set to 0, allowing for automatic adjustments to ensure a balance between determinism and diversity. We then generated 10 texts for the positive category, repeating this 100 times to produce 1,000 synthetic texts for both Tasks 5 and 6. The schematic diagram of the prompts is shown in the figure 1. Additionally, the examples of the full prompts and the examples of the generated texts are presented in the appendix A.1 and A.2.

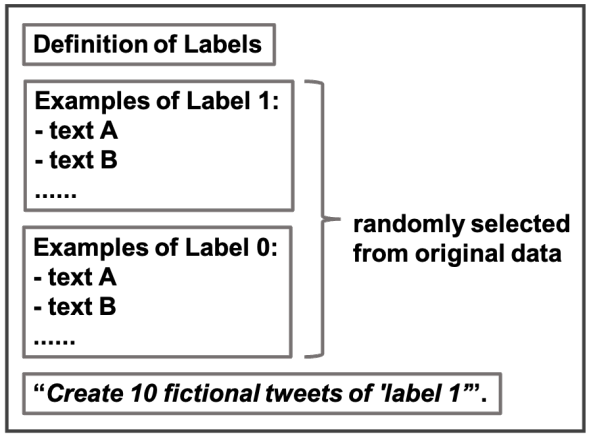

Figure 1: Schematic diagram of the prompt. Composed of the definition of labels, specific examples of texts, and instructions for generation.

3.2 Model Development

For the development of our text classification models, we utilized the DeBERTa v3 Large model (He et al., 2021b,a). For each task, we trained two versions of the model: one using synthetic data for training and the other without using synthetic data.

3.2.1 Training Procedure

Training was conducted over 10 epochs, starting with a learning rate of 5e-5. A scheduler was used to reduce the learning rate to zero towards the end of the training process.

3.2.2 Validation and Model Selection

The model's performance was evaluated on the validation dataset at the end of each epoch using the F1-score. The model that achieved the highest F1-score on the validation set was selected for inference on the test dataset.

3.3 LLM-Generated Text Assessment

3.3.1 Clustering of Texts Embeddings

In the text analysis using sentence transformers, we extracted embeddings using "all-MiniLM-L6-v2" available at https://github.com/UKPLab/sentence-transformers (Reimers and Gurevych, 2019). On top of this, we performed k-means clustering and t-distributed stochastic neighbor embedding (t-SNE) to visualize and evaluate the distribution of the synthetic data and the raw data.

3.3.2 *N*-gram Analysis

Additionally, we conducted the same n-gram analysis for ($n = 1, 2, 3$) on the raw data of Tasks 5 and 6 for comparison and discussion. To ensure a fair comparison with the GPT-4 generated text, which had a limited number of samples, we randomly selected 1,000 samples from each of the original datasets (Label 0 and Label 1) for both tasks.

4 Results

4.1 Performance Metrics on Test Data

The evaluation metrics for the test data of Tasks 5 and 6 are presented in Table 2. For the primary metric, the F1 score, Task 5 shows a slight improvement of 0.009, while Task 6 shows a deterioration of 0.013.

Table 2: Performance metrics for Tasks 5 and 6 with and without data augmentation using LLM-generated texts (indicated by "Aug"). **Bold** indicates the higher score for each task.

Task	Aug	F1	Precision	Recall
5	No	0.924	**0.966**	0.886
	Yes	**0.933**	0.932	**0.934**
6	No	**0.936**	**0.947**	**0.926**
	Yes	0.923	0.921	0.924

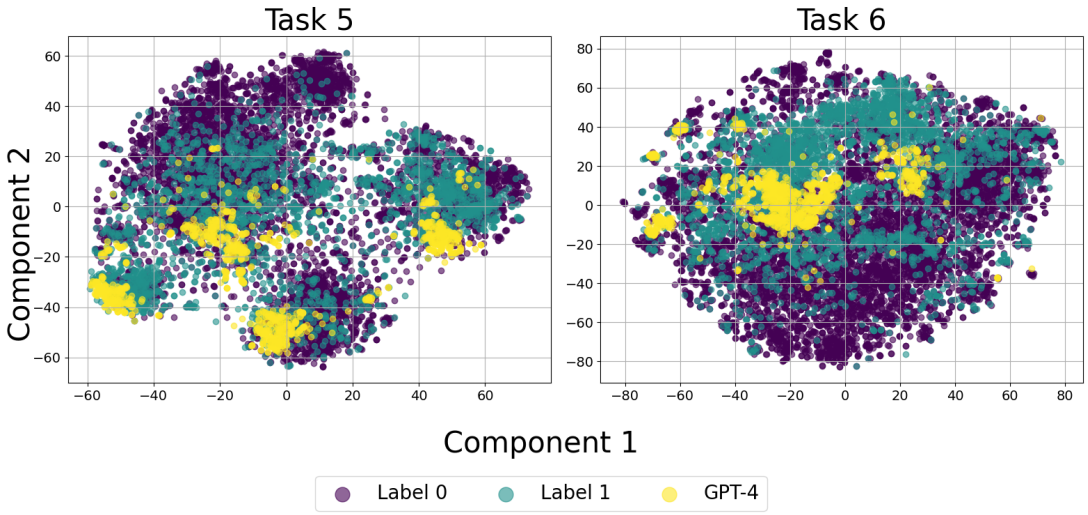

Figure 2: t-SNE visualization of sentence embeddings for Tasks 5 and 6. The plots show the distribution of original and GPT-4 generated sentences in a two-dimensional space after dimensionality reduction using t-SNE. Each point represents a sentence, with colors indicating the label (Label 0, Label 1, or GPT-4).

4.2 *N*-gram Analysis

Based on the results presented in Table 3, it is evident that the texts generated by GPT-4 consistently exhibits a lower number of unique n-grams compared to the original dataset, across both Tasks 5 and 6. This observation holds true for all values of n (1, 2, and 3) considered in the analysis.

Table 3: Number of unique n-grams for each label and task. "Label 0" and "Label 1" represent data from the original dataset, while "GPT-4" represents text generated by GPT-4.

Task	Data	n=1	n=2	n=3
5	Label 0	5,380	22,313	30,577
	Label 1	5,001	21,056	29,102
	GPT-4	2,522	12,922	19,386
6	Label 0	5,471	15,776	18,601
	Label 1	3,337	10,696	13,186
	GPT-4	1,640	6,940	10,331

4.3 Clustering of Text Embeddings

Figure 2 illustrates that GPT-4-generated text embeddings form localized clusters with limited spread compared to the original data, particularly Label 0 sentences. This suggests a lack of diversity in GPT-4 outputs, as the model tends to generate semantically similar sentences, in contrast to the wider distribution and linguistic variety observed in human-generated text.

5 Discussion & Conclusion

The data provided for the Tasks 5 and 6 were both imbalanced datasets with less positive examples than negative ones. To address such imbalances, data augmentation can be an effective approach. Previous methods using deep learning have been proposed as techniques applicable to named entity recoginition and text classification, such as label-wise token replacement, synonym replacement, and entity replacement within sequences (Dai and Adel, 2020; Ding et al., 2021; Zhou et al., 2022). There have also been proposals for approaches combining context and entity levels using LLMs (Ye et al., 2024). In this study, we proposed a method that employs few-shot learning techniques to generate new text using LLMs (Brown et al., 2020). By leveraging LLMs, a vast amount of completely new data can be generated, and by producing high-quality data, improvements in the performance of classification models can be expected. However, in this study, the addition of synthetic data did not significantly improve the performance compared to the baseline model.

As evident from the visualization by clustering, the synthetic data generated by GPT-4 for the Tasks 5 and 6 exhibits a localized distribution in the embeddings extracted from the pre-trained language model, compared to the original data. This suggests that the synthetic data lacks diversity in comparison to the original data, and as a result, the addition of the synthetic data did not improve the overall diversity of the training data, resulting in no explicit

improvement in the model's performance.

This study has several limitations. Since we used only GPT-4 for generating synthetic data from sources like Reddit and Twitter, future work should explore new methods to increase diversity. Prior research suggests that diversity can be enhanced by specifying attributes (Yu et al., 2024). For instance, identifying the characteristics of posters or the types of children's diseases could lead to greater diversity. Future efforts could include using multiple language models, presenting more original examples, or adjusting hyperparameters like temperature to improve data diversity.

The Tasks 5 and 6 involved relatively few positive examples, resulting in imbalanced data, which is often the case in real-world settings. While the use of synthetic data generated by language models can be an effective solution for augmenting training data, this study suggests that various efforts are necessary to create diverse data that is effective for training language models.

Acknowledgments

We would like to acknowledge that all ethical aspects of this research were strictly adhered to, in accordance with the regulations set forth by the data providers. Additionally, this study was conducted without the use of any research funding.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. 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)*, pages 3198–3213, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Zahra Shakeri Hossein Abad, Adrienne Kline, Madeena Sultana, Mohammad Noaeen, Elvira Nurmambetova, Filipe Lucini, Majed Al-Jefri, and Joon Lee. 2021. Digital public health surveillance: a systematic scoping review. *NPJ digital medicine*, 4(1):41.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health applications (smm4h) shared tasks at acl 2024. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Full Prompts

The full texts of the prompts are provided below. The variables `pos_exs` and `neg_exs` each contain five example texts for Labels 1 and 0, respectively.

A.1.1 Prompt for Task 5

```
The following tweets are from a
   parent with a child.
The label has a definition like '
   This binary classification
   task involves automatically
   distinguishing tweets, posted
   by users who had reported
   their pregnancy on Twitter,
   that report having a child
   with attention-deficit/
   hyperactivity disorder (ADHD),
    autism spectrum disorders (
   ASD), delayed speech, or
   asthma (annotated as "1"),
   from tweets that merely
   mention a disorder (annotated
   as "0").

Examples of label 0:
{neg_exs}

Examples of label 1:
{pos_exs}

Create 10 fictional tweets of '
   label 1'.

You should output following style
   .
1. output1
2. output2
3. output3
4. output4
5. output5
6. output6
7. output7
8. output8
9. output9
10. output10
```

A.1.2 Prompt for Task 6

```
The posts in the following
   dataset come from social media
    platforms Twitter and Reddit.
    These posts are used to
   automatically train models
   focused on identifying
   instances where the user's
   exact age is explicitly
   mentioned. This is
   particularly useful for health
    and demographic research
   applications that require
   precise age data.

Heres a summary of the
   definitions for labels 0 and 1
    in this dataset:

Label 1: This label is assigned
   to posts where the user's
   exact age can be determined
   directly from the text at the
   time the entry was posted.
   Examples include explicit
   mentions of age, such as "It's
    my 21st birthday today" or
   inferred statements where the
   user indicates they will be a
   certain age, like "tomorrow I'
   ll be 20."
Label 0: This label is used for
   posts where the age of the
   user cannot be determined or
   is ambiguous. Examples include
    unclear references to age,
   mentions of age that may not
   be current (like past or
   future tense without a clear
   indicator of current age), or
   mentions of someone elses
   age (e.g., a sibling or child)
   .

Examples of label 1:
{pos_exs}

Examples of label 0:
{neg_exs}

Create 10 fictional tweets of '
   label 1'.

You should output following style
   .
1. output1
2. output2
3. output3
4. output4
5. output5
```

```
6. output6
7. output7
8. output8
9. output9
10. output10
```

A.2 Examples of Generated Texts

Here are three actual generated examples from the 1,000 texts produced by GPT-4 for each of the two tasks.

A.2.1 Examples of Task 5

1. Just had a parent-teacher conference about my daughter's ADHD. The teacher recommended some strategies to help her stay focused in class. Feeling hopeful and supported. #ADHDawareness

2. Navigating ADHD with my child has been a journey of patience, love, and a lot of learning. But seeing his improvements makes it all worth it. #ParentingADHD

3. My toddler with a speech delay said "mama" clear as day. I cried. These moments are everything. (pleading face emoji)(heart with arrow emoji) #speechdelay

A.2.2 Examples of Task 6

1. Just hit the big 25 today, can't believe I'm a quarter of a century old! (party popper emoji)(birthday cake emoji) #birthdayvibes

2. Just signed the lease to my very first apartment, a perfect 27th birthday present to myself. Here's to independence and new beginnings! (house emoji)(birthday cake emoji) #NewHome #27Years

3. Turning 22 in a pandemic means virtual birthday parties and lots of Zoom shots! #QuarantineBirthday