Assessing the Efficacy of Pre-Trained and Large Language Models for Health Classification with Varying Data Volumes

Anonymous ACL submission

Abstract

Automated text classification in medical and health domains enables the extraction of structured information from unstructured clinical text, such as identifying diseases and associated conditions. However, applying text classification models effectively in healthcare requires a nuanced understanding of specific subtopics and the trade-offs between model scale and available data resources. This paper evaluates the performance of pretrained language models (PLMs) and large language models (LLMs) in classifying subtopics within the sleep and activity domains. Using a dataset of curated Reddit posts, we examine how classifier performance varies with different training sample sizes, including low-resource scenarios with just one to five examples. Our findings highlight a complex interaction between model architecture, data availability, and classification performance, demonstrating the strengths of LLMs in zero-shot learning in nuanced subdomains with limited data, while PLMs surpass LLMs with modest increases in data. This research provides valuable insights into the optimal application of language models for health-related text classification tasks, especially under varying resource constraints.

1 Introduction

The utilization of large language models (LLMs) in personal health has garnered significant attention, particularly due to their potential in generating personalized health recommendations and processing health-related data (Huang et al., 2019; Thirunavukarasu et al., 2023; Singhal et al., 2023; Yang et al., 2022). Recent research reveals a critical need for models capable of handling the nuanced and context-dependent nature of health information (Guo et al., 2024; Harris et al., 2024). Advancements in transformer architectures, particularly encoder models like BERT (Devlin, 2018) and decoder models like the Generative Pre-trained Transformer (GPT) series (Radford, 2018; Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023), have fueled progress in this area (Guo et al., 2024; Harris et al., 2024). However, prior research involving BERT models and their variants typically explores thousands of examples per class, which may require substantial data labeling efforts to train these models (Guo et al., 2024; Mujtaba et al., 2019). LLMs, such as GPT series, with their extensive pre-training, offer a potential solution by reducing reliance on labeled data. On the other hand, while LLMs have shown promise in broad health domain classifications, particularly when compared to pre-trained learning models (PLMs) trained on extensive datasets (e.g., >5000 examples (Guo et al., 2024)), their ability to discern subtle distinctions within related health subdomain areas remains less explored. This highlights the need for models that can accurately identify subdomains within specific primary, parent health domains, and an understanding of how model scale and data resources influence this capability.

To investigate this, we evaluate the performance of both large language models (LLMs) and state-of-the-art text classification techniques on a dataset encompassing sleep and fitness-related text. We test 4 encoder-based PLMs: BERT-base (Devlin, 2018), DistilBERT (Sanh, 2019), Electra-base, and Electra-small ("Efficiently Learning an Encoder that Classifies Token Replacements Accurately") (Clark, 2020). We contrast them against 2 Large Language Models in the Gemini family - Nano and Pro (Anil et al., 2023). Specifically, this study contributes: (1) a comparative analysis of PLMs and LLMs on distinct (primary) health domain and respective subdomains classification; and (2) an evaluation of model stability using a synthetic dataset. Furthermore, recognizing the practical constraints often encountered in real-world healthcare settings, (3) this work explores performance across a spectrum of resource conditions, from extremely lim-

ited training data (1-50 examples per class) to the larger datasets characteristic of prior work, providing valuable insights into the optimal deployment of LLMs and PLMs in diverse healthcare contexts. This investigation provides valuable insights for optimizing the deployment of LLMs and PLMs in diverse healthcare applications and advances our understanding of how these models can be used to effectively parse and classify complex health information.

2 Related Work

The field of medical text classification has seen extensive exploration, leveraging both traditional machine learning and advanced deep learning techniques (Adeva et al., 2014; Mujtaba et al., 2019; Qing et al., 2019; Hughes et al., 2017; Lu et al., 2022). Automated text classification can enable a deeper understanding and can support medical experts in processing health text, including disease status, laboratory results, medication history, side effects, and treatment outcomes (Harris et al., 2024; Adeva et al., 2014; Mujtaba et al., 2019). Several studies have explored the development of domain-specific models by pre-training or fine-tuning models on specialized corpora to enhance their efficacy in specific areas. BERT (Bidirectional Encoder Representations from Transformers) and its variants have become state-of-the-art models in medical NLP (Kim et al., 2023; Khadhraoui et al., 2022; Wang et al., 2021). For example, Kim et al. (2023) proposed a model using a domain-specific pre-trained BERT (KM-BERT) to predict medical specialties from patient-provided question text, achieving improved performance compared to other deep learning NLP models and demonstrating its potential to benefit hospital patient management, Wang et al. (2021) proposes a medical triage system using BERT to classify patient symptom descriptions into medical specialties, aiming to reduce hospital triage pressure and achieved relatively high accuracy in classifying patient questions, demonstrating the potential to help patients choose appropriate consultation rooms and alleviate the triage burden in hospitals. By training on specialized corpora, these models can better understand the unique language, terminology, and relationships within their respective domains, however, Mujtaba et al. (2019) in a review of clinical text classification research trends highlights that while many machine learning techniques are effective in medical text classification tasks, they require substantial human effort to create labeled training data often requiring thousands of examples per class. This dependency on labeled data becomes a significant hurdle, especially for specialized or emerging domains, and may result in imbalanced dataset (Prabhakar and Won, 2021; Lu et al., 2022).

Large language models (LLMs) have emerged as potential zero-shot classifiers for various tasks, including health topic classification, leveraging their extensive pre-trained knowledge to minimize reliance on labeled data (Harris et al., 2024). However, studies have yielded mixed results regarding the effectiveness of LLMs in this role compared to supervised models. For example, Harris et al. (2024) assessed several LLMs (e.g., GPT-4, Llama-3-8B) across 17 public health tasks and observed varied performance: LLMs performed well on simpler tasks like gastrointestinal illness classification, but struggled with more domain-specific tasks such as virology or contact type classification. Similarly, Guo et al. (2024) found that supervised PLMs, such as RoBERTa, BERTweet, and SocBERT, generally outperformed LLMs in classifying health information across six social media-based health topics. This underscores the value of PLMs when trained on task-specific data. However, LLMs did achieve near-human performance in identifying self-reported depression, suggesting their potential effectiveness in certain contexts that require nuanced language or subjective experience understanding.

3 Datasets and Models

Below, we detail the dataset creation used in experiments including data gathering, pre-processing, labeling, as well as model selection processes.

3.1 Gathering Real-World Personal Health Scenarios

This section details the creation and preparation of datasets used to evaluate the performance of various classification models on English health-related social media text. The data for this study was sourced from publicly available, archived Reddit posts spanning from June 2005 to December 2022, hosted on The Eye, an open library for the Pushshift Reddit Dataset (Baumgartner et al., 2020)[1]. No user IDs and author names were included in extracted datasets. A total of

[1]https://the-eye.eu/redarcs/

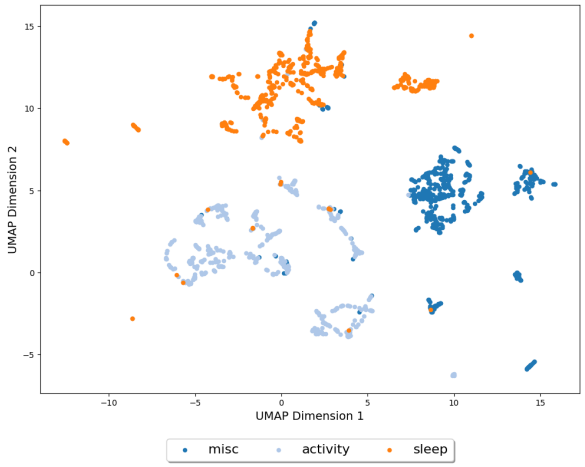

Figure 1: UMAP Visualization of BERT Embeddings for a Sample of Primary Classes (n=500 per class)

12 subreddits related to sleep (r/sleep, r/insomnia, r/apnea, r/sleepparalysis, r/narcolepsy) and fitness (r/exercise, r/fitness, r/running, r/cycling, r/swimming, r/hiking, r/weightlifting) were selected. Each post underwent cleaning processes, including lemmatization, stop word removal, and URL removal, to ensure data quality.

Posts were initially categorized into "Sleep" and "Activity" datasets based on their originating subreddit. To further refine these datasets into nuanced subdomains, BERTopic (Grootendorst, 2022), a topic modeling technique leveraging BERT embeddings, was employed. A custom Maximal Marginal Relevance representation model, with a diversity parameter of 0.2, was integrated within the BERTopic framework. Additional parameters included calculate_probabilities=False, verbose=True, n_gram_range=(1, 2), and min_topic_size=50 to generate diverse and meaningful topics with a minimum size threshold.

While BERTopic dynamically generated topics, 20 subdomains were manually curated for the Sleep dataset and 21 for the Activity dataset. Each subdomain contained a minimum of 100 posts, with an additional "miscellaneous" category encompassing posts not fitting the curated subdomains. This manual curation ensured balanced representation and manageable label sets for analysis. The decision to limit the number of categories was driven by a desire to create a manageable and meaningful set of labels. A larger number of categories would have likely resulted in overlapping topics and decreased clarity in analysis. Posts with multiple topics were split, and to visualize differences between the BERT sentence embeddings, we employed UMAP (Uniform Manifold Approximation and Projection), a dimensionality reduction technique for visualization (McInnes et al., 2018), on a random sample of 500 posts per primary class (Figure 1) and 50 posts per subdomain (Figures 5 and 6 in the Appendix). This visualization, along with semantic similarity search using a cosine similarity threshold of 0.8, helped identify and resolve posts with overlapping subdomains via manual relabelling or removal, further refining the datasets and ensuring distinct category separation. This resulted in the creation of the following new datasets:

Sleep Dataset. The resulting Sleep dataset contains 2484 posts with subdomains such as alarm, bed comfort, beverage effects, and sleep movement (c.f. Appendix 1 for a full list and descriptions).

Activity Dataset. The Activity dataset consists of 2159 posts with subdomains like heart health, lower body exercise, abdominal exercise, and upper body exercise (c.f. Appendix 2).

Out-of-distribution Dataset. An out-of-distribution dataset was also created, comprising 2348 unique posts randomly sampled from 10 non-sleep and non-activity related subreddits (e.g., classical music, education, pets, relationships).

Synthetic Dataset To create realistic online discussion posts for subdomain classification within the Sleep and Activity categories, we used Gemini-pro. For each subdomain, Gemini-pro generated 20 posts per class based on prompts related to that specific topic (e.g., "Generate an online post about alarms, alarm clocks, or waking up with alarms."). Our primary domain classification task used 180 posts: 60 each related to activity, sleep, and a set of 60 out-of-distribution posts. The out-of-distribution set was created by generating posts from prompts that excluded any subdomain-specific keywords or concepts.

Models We evaluated several models, including BERT-base (Devlin, 2018), DistilBERT (Sanh, 2019), ELECTRA-base and small (Clark, 2020), against the Gemini family of models (Pro and Nano) (Anil et al., 2023) for each classification task (cf. Appendix A.1 for more model details and A.2 for computation cost details).

3.2 Analysis

To assess model performance across diverse training data sizes, we conducted experiments on two classification tasks: a primary task (sleep, activity, out-of-distribution) and a subdomain task (specific topics within sleep and activity). BERT and ELEC-

TRA models were fine-tuned on varying training set sizes, ranging from a small number of examples per class (1, 2, 5, 10, 50, 100, 150, 1000, and the full real-world testing dataset) using the Hugging Face Transformers library (Wolf, 2019). Gemini models, in contrast, were evaluated in a few-shot learning setting, utilizing zero-shot prompting with an option to provide up to 10 examples (due to context window length limitations) per class during inference. Because model performance can be influenced by the input prompt (Liu and Shi, 2024), we used a consistent template across all classification tasks, with only keywords and descriptions modified to reflect the specific task. This prompt template was iteratively refined until stable performance was observed across multiple runs. Models were instructed to return a single class label. Inputs without a recognized label were initially designated as "Unknown," though this category was not observed in practice during experimentation. Performance was measured using accuracy as the manually curated dataset had minimal class imbalance. The same procedures were done on the synthetic dataset. Given the potential variability in performance across runs for LLMs like the GPT series (Harris et al., 2024), each task was repeated three times for the Gemini models, with the average performance reported. The variance across runs for both real-world and synthetic dataset was minimal (0.0-0.1), indicating relatively stable performance (cf. Appendix 3 and 4 for more details).

4 Findings

4.1 Model Performance and Data Volume For the Primary Domain Classification Task

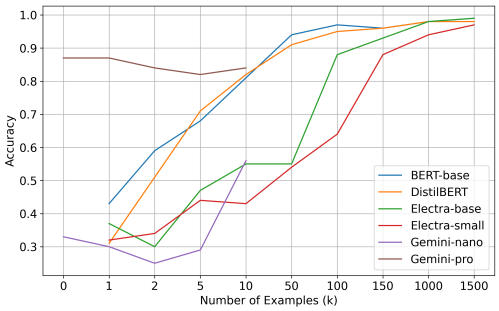

Figure 2: Model Performance on Primary Domain Classification at Different k Values.

In the three-class classification task (sleep, activity, out-of-distribution) as shown in Figure 2, more training examples generally translated to higher accuracy for most models. However, the rate of improvement and the performance ceiling varied significantly. The BERT models achieved high accuracy (90%) with as few as 50 examples, indicating their aptitude for learning generalizable patterns from limited data. ELECTRA-base, compared to ELECTRA-small, exhibited faster learning and a higher performance ceiling, likely due to its increased capacity. However, both ELECTRA variants ultimately achieved comparable accuracy to the other BERT models, albeit with slower learning curves, generally requiring more examples to reach peak performance. Gemini-nano, constrained by its smaller context window and limited capacity for processing large examples per class, also improved with additional data, eventually exceeding or matching the ELECTRA models' performance on smaller datasets (k=10). However, it lagged considerably behind BERT models with larger datasets, highlighting the limitations of its smaller scale.

In contrast, Gemini-pro exhibited a relatively flat learning curve, achieving 80% accuracy irrespective of data volume. Gemini-pro, the largest model tested, achieved the highest accuracy with minimal examples. However, its performance plateaued with increasing data. This suggests a strong reliance on pre-existing knowledge, requiring minimal task-specific tuning, yet struggling to capture finer nuances that benefited smaller BERT models as datasets grew. Notably, with sufficient training data (k$\geq$ 50), the BERT and ELECTRA models surpassed Gemini-pro's performance obtained with the smaller datasets.

4.2 Model Performance and Data Volume For the Subdomain Classification Task

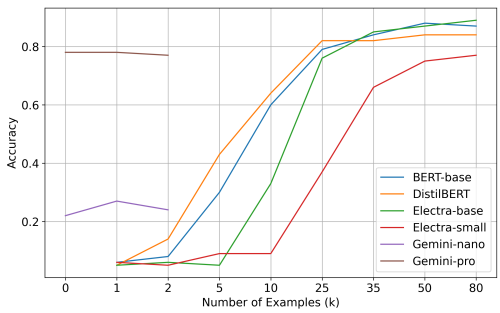

Figure 3: Model Performance on Sleep Subdomain Classification at Different k Values

A similar pattern emerged for the more granular

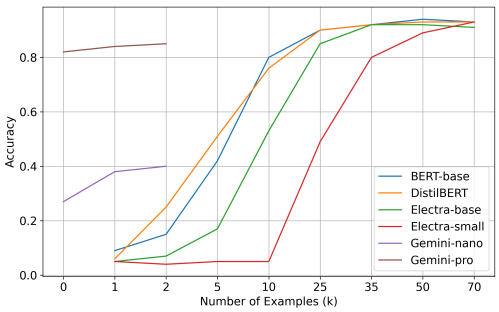

Figure 4: Model Performance on Activity Subdomain Classification at Different k Values

21-subdomain classifications (for both sleep and activity) as seen in Figures 3 and 4. The BERT and ELECTRA models again benefited from increased data, with performance steadily improving as the number of examples grew. Interestingly, these models achieved high accuracy levels (>80% for sleep and >90% for activity) with roughly the same or less number of examples ($\geq$ 50) as in the three-class task. This might suggest that, for these models, the increased complexity of the 21-subdomain task was offset by the richer information content inherent in the more specific labels. These models also required slightly more examples for sleep sub-classification than activity sub-classification likely due to classes being more heterogenous for activity and primary datasets. Likely due to the increased data requirements for nuanced classification and its limited context window, Gemini-nano did not exhibit clear performance gains with additional data in this setting; even fewer examples could fit within its context window when numerous subdomains were present, hindering its ability to learn effectively.

On the other hand, Gemini-pro once again displayed remarkable generalization capabilities, achieving high accuracy (>77% for sleep, >82% for activity) with minimal data (as few as 2 examples). This reinforces the trade off between strength in leveraging pre-existing knowledge for specialized classification within the health domain, even when confronted with limited task-specific examples, and limited abilities to capture the subtleties of nuances of data not reflected in the training set. Interestingly, for smaller training datasets (zero to few examples), both Gemini models, including Gemini-nano, outperformed the BERT and ELECTRA models. This suggests that the inherent knowledge encoded in the Gemini-pro models, even the smaller Gemini-nano, provided an initial advantage when task-specific data was scarce. However, this advantage diminished quickly as the BERT and ELECTRA models rapidly improved with the addition of just a few more examples per class.

4.3 Model Performance on the Synthetic Dataset

We replicated our analysis on a synthetic dataset to validate our findings and assess potential overfitting to the Reddit data. The overall trends remained consistent, confirming the robustness of our observations. BERT models efficiently achieved high accuracy on all classification tasks with a modest number of examples, with ELECTRA models required more data. While Gemini-nano was initially competitive, its performance fell behind BERT models and ELECTRA-base as training data increased. Gemini-pro, once again exhibiting a flat learning curve, consistently outperformed all other models, likely due to its inherent advantage in recognizing patterns within the synthetic data it generated. Nevertheless, with sufficient training, BERT models reached comparable performance levels. This consistency across datasets indicates that the observed patterns are not solely artifacts of the original Reddit dataset.

5 Discussion and Conclusion

5.1 Implications for Health Tasks

Our findings reveal a nuanced interplay between model architecture, data volume, and performance in health-related text classification, with important implications for practical applications. Consistent with the observations of Guo et al. (2024) in broader health domains, we identify a trade-off between leveraging pre-existing knowledge (LLMs) and data-driven learning (PLMs). This trade-off is evident not only across different health domains but also within more specialized subdomains. Specifically, in scenarios where data is limited and the classification task involves multiple nuanced subdomains—such as those encountered in subjective health experiences like sleep and activity—larger LLMs, like Gemini-pro, tend to perform better. These models, with their extensive pre-training, possess a wealth of knowledge that allows them to generalize effectively from relatively few task-specific examples, thus outperforming PLM models in data-constrained situations. This observation aligns with Guo et al. (2024)'s findings, which

highlight GPT-4's strong performance on the complex task of depression detection. This suggests that LLMs, including the Gemini family explored in our study, may be particularly well-suited for classifying subjective experiences in health contexts. The superior performance of larger LLMs emphasizes the importance of model size, particularly when utilizing pre-existing knowledge to overcome limited task-specific data. In contrast, smaller models like Gemini-nano, which are designed for on-device use, struggle to generalize effectively in zero-shot settings due to their limited pre-trained knowledge and parameter capacity. While the compact design of Gemini-nano makes it ideal for mobile health applications, where device constraints demand smaller models, its performance in data-scarce environments highlights the challenges of deploying such models for complex health tasks. This trade-off between size and capability is critical when considering the practicality of mobile health applications, where on-device processing often takes precedence.

However, the strong zero-shot performance of larger LLMs comes with a significant computational cost. For example, Gemini-pro required a substantial number of GPU hours—132 hours—compared to PLMs like BERT and ELECTRA, which only required 6-8 hours (see Appendix A.2). This disparity becomes even more noticeable when considering smaller LLMs, such as Gemini-nano, which, despite not consistently demonstrating strong performance in zero-shot settings, still required more training time (approximately 11 hours) than PLMs. Additionally, PLMs like BERT and ELECTRA not only trained faster but also showed considerable improvement with even minimal additional data, performing well on both primary classification tasks and more complex subdomain tasks. In some cases, they eventually surpassed LLMs with limited examples. This supports the findings of Guo et al. (2024), who showed that supervised PLMs outperform zero-shot LLMs in health classification tasks when sufficient data is available. In contrast to the thousands of examples per class typically explored in prior research, our study extends these findings by showing that this effect holds true even with as few as 50 examples per class, even as tasks include subjective experiences, become more complex, and the number of classes increases. The consistent performance patterns observed across both the real-world Reddit dataset and a synthetic dataset further validate these results, indicating that these trends reflect broader patterns in model behavior based on size. Thus, the choice between LLMs and PLMs becomes a critical consideration, requiring a careful balance between performance goals, available computational resources, and data volume.

5.2 Limitations and Future Work

While this study provides valuable insights, it acknowledges the rapid pace of development in large language models (LLMs). The evaluation was conducted using specific model architectures and focused on the sleep and fitness domains, which may limit the generalizability of findings to other health areas and more recent models (Warner et al., 2024). Notably, advancements like the Gemini family of models have emerged since the study's inception, potentially offering improved performance or efficiency. Future research should explore the generalizability of these results across diverse health domains and updated model architectures. This includes exploring innovative training approaches that leverage task-specific data and examining hybrid architectures that combine the strengths of LLMs and encoder-based PLMs like BERT. For instance, ensembling methods, such as those employed by Zhou et al. (2023) for clinical note classification, could leverage the complementary capabilities of different model types. This could be particularly relevant for resource-constrained environments like mobile devices, where a locally-run PLM and small LLM combination could offer comparable performance to a large LLM while enhancing privacy by eliminating the need for remote server processing and data transmission.

Additionally, while the synthetic dataset was intended to investigate models overfitting to the Reddit data, it may not fully capture the nuances and complexities of real-world health information. Future studies should incorporate a wider variety of datasets and evaluation metrics to develop a more comprehensive understanding of the strengths and weaknesses of various language models for health-related text classification.

5.3 Risks

This research focuses on exploring and evaluating the capabilities of various language models for specific tasks within the sleep and activity domains. Although this study does not involve deployment or direct application in a clinical setting, the findings contribute to the growing body of knowledge re-

garding the use of language models in health. The primary risks in our work are related to internal model and training set biases, which may reflect historical inequities in healthcare. These biases could inadvertently influence model predictions, potentially leading to unfair or discriminatory outcomes if applied in real-world setting, highlighting the need for careful data curation before actual deployment.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

JJ García Adeva, JM Pikatza Atxa, M Ubeda Carrillo, and E Ansuategi Zengotitabengoa. 2014. Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, 41(4):1498–1508.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 1.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

K Clark. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Yuting Guo, Anthony Ovadje, Mohammed Ali Al-Garadi, and Abeed Sarker. 2024. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, 31(10):2181–2189.

Joshua Harris, Timothy Laurence, Leo Loman, Fan Grayson, Toby Nonnenmacher, Harry Long, Loes WalsGriffith, Amy Douglas, Holly Fountain, Stelios Georgiou, et al. 2024. Evaluating large language models for public health classification and extraction tasks. *arXiv preprint arXiv:2405.14766*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Mark Hughes, Irene Li, Spyros Kotoulas, and Toyotaro Suzumura. 2017. Medical text classification using convolutional neural networks. In *Informatics for health: connected citizen-led wellness and population health*, pages 246–250. IOS Press.

Mayara Khadhraoui, Hatem Bellaaj, Mehdi Ben Ammar, Habib Hamam, and Mohamed Jmaiel. 2022. Survey of bert-base models for scientific text classification: Covid-19 case study. *Applied Sciences*, 12(6):2891.

Yoojoong Kim, Jong-Ho Kim, Young-Min Kim, Sanghoun Song, and Hyung Joon Joo. 2023. Predicting medical specialty from text based on a domain-specific pre-trained bert. *International Journal of Medical Informatics*, 170:104956.

Menglin Liu and Ge Shi. 2024. Poliprompt: A high-performance cost-effective llm-based text classification framework for political science. *arXiv preprint arXiv:2409.01466*.

Hongxia Lu, Louis Ehwerhemuepha, and Cyril Rakovski. 2022. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC medical research methodology*, 22(1):181.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Ghulam Mujtaba, Liyana Shuib, Norisma Idris, Wai Lam Hoo, Ram Gopal Raj, Kamran Khowaja, Khairunisa Shaikh, and Henry Friday Nweke. 2019. Clinical text classification research trends: systematic literature review and open issues. *Expert systems with applications*, 116:494–520.

Sunil Kumar Prabhakar and Dong-Ok Won. 2021. Medical text classification using hybrid deep learning models with multihead attention. *Computational intelligence and neuroscience*, 2021(1):9425655.

Li Qing, Weng Linhong, and Ding Xuehai. 2019. A novel neural network-based method for medical text classification. *Future Internet*, 11(12):255.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Xinyuan Wang, Make Tao, Runpu Wang, and Likui Zhang. 2021. Reduce the medical burden: An automatic medical triage system using text classification bert based on transformer structure. In *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pages 679–685. IEEE.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Weipeng Zhou, Dmitriy Dligach, Majid Afshar, Yanjun Gao, and Timothy A Miller. 2023. Improving the transferability of clinical note section classification models with bert and large language model ensembles. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 125. NIH Public Access.

A Appendix

A.1 Detailed Model Descriptions

We evaluated several models for each classification task. These include:

- **BERT-base:** This model has 12 layers, 110M parameters, and a hidden dimension size of 768 (Devlin, 2018).
- **DistilBERT:** This model is a distilled version of BERT, and has 66 million parameters (Sanh, 2019).
- **ELECTRA-Base:** This model has 12 layers, 110M parameters, and a hidden dimension size of 768 (Clark, 2020).
- **ELECTRA-Small:** This model has 12 layers, 14M parameters, and a hidden dimension size of 256. (Clark, 2020).
- **Gemini Pro:** This is a performance-optimized model in terms of both cost and latency (Anil et al., 2023).
- **Gemini Nano:** This model is distilled from larger Gemini models, and has 1.8B parameters (Anil et al., 2023).

A.2 Model and Experimental Setup Details

Using a V100 GPU, the DistilBERT and ELECTRA-small models required around 6 hours of GPU time each, Bert-base and ELECTRA-base required around 8 hours, Gemini-nano required around 11 hours, and Gemini-pro required around 132 hours, utilizing a total of approximately 171 GPU hours across different models.

Optimal hyperparameters, including learning rate and batch size, were determined using a validation set. The best performing BERT-base models were trained with a learning rate of 3e-5 and a batch size of 16. DistilBERT, ELECTRA-Base, and ELECTRA-Small achieved optimal performance with a batch size of 16 and learning rates of 2e-5, 3e-5, and 5e-6, respectively. All models utilized a maximum sequence length of 512. Early stopping was implemented based on validation accuracy. Sampling ensured representation of all unique labels in each training subset. For the Gemini models, we employed their default configurations and set the sampling temperature to 0.0 to encourage reproducibility.

A.3 Analysis Details

Datasets for both classification tasks were prepared from a combination of real-world and synthetic sources. The "out-of-distribution" domain was exclusively used in primary classification. For BERT and ELECTRA, training data sizes ranged from 1

to 1000 examples per class (and the full dataset), while for Gemini, sizes ranged from zero-shot to 10 examples per class. For a more granular micro analysis, we used training sizes from zero-shot to 50 examples per class for BERT and ELECTRA and zero-shot to 2 examples per class for Gemini. Data was stratified into training, validation, and test sets. Textual labels were encoded numerically, and an "unknown" category was included for unseen labels (though unused in practice for Gemini).

A.4 Subdomain Details and Descriptions

Table 1 and 2 detail the subdomains for sleep and activity datasets, along with descriptions for each subdomain.

A.5 UMAP Visualizations of Activity and Sleep Subdomains

Figures 6 and 5 present UMAP visualizations of BERT embeddings for samples from the sleep and activity subdomains, respectively. These visualizations reveal both the proximity between subdomains and their distinct separations.

A.6 Complete Results

Tables 3 and 4 present the classification accuracy for each model across various support examples provided during training or few-shot learning. The Gemini models' performance includes standard deviation from 3 runs.

A.7 Prompts

Table 5 presents the prompt templates used for different text classification tasks leveraging a Large Language Model (LLM). The table is organized by task, with each row corresponding to a specific classification objective: primary topic classification (into 'sleep', 'activity', and 'out-of-distribution') and sub-classification within the 'activity' and 'sleep' categories. For the primary classification task, both zero-shot (no examples provided) and few-shot (examples provided) prompt variations are detailed. The prompts for sub-classification tasks encompass both zero-shot and few-shot scenarios, where the model is given detailed descriptions of each subcategory. All prompts instruct the LLM to act as a topic classifier and to return a single label corresponding to the most appropriate category for the input post. The placeholders {post}, {examples[i]} and {labels[i]} indicate where the input post, example posts, and corresponding labels are dynamically inserted into the prompt, respectively. Placeholder {Description for each activity/sleep category} indicates where detailed descriptions of each subcategory are inserted.

Table 1: Sleep Dataset Subdomains and Descriptions

Label	Includes discussion around
Alarm	alarms, alarm clocks, or waking up using alarms
Bed Comfort	about bed comfort, mattress type, pillows, etc.
Beverage Effects	effects of beverages (e.g., coffee, tea) on sleep
Sleep Movement	sleepwalking, sleep talking, or other movement during sleep
Cannabis	use of cannabis and its effects on sleep
Cataplexy	sudden loss of muscle control, often associated with narcolepsy
Diet	the relationship between diet and sleep
Dream	dreams, vivid dreams, or dream interpretation
Exercise	the relationship between exercise and sleep
Gendered Health	the relationship between menstruation, pregnancy, or other gendered factors and sleep
Light Source	the impact of light sources (e.g., sunlight, artificial light) on sleep
Mental Health	the relationship between mental health and sleep
Nap	napping, nap schedules, or nap strategies
Nightmare	nightmares, their frequency, or coping mechanisms
Sleep Position	preferred sleep positions or their impact on sleep quality
Sleep Schedule	sleep schedules, routines, or sleep hygiene
Sound Control	using white noise, earplugs, or other methods to control sleep environment sounds
Supplements	using supplements (e.g., melatonin, valerian) to improve sleep
Temperature	the impact of room temperature on sleep quality
Tracker	using sleep trackers or wearable devices to monitor sleep
Miscellaneous	any sleep-related topic not covered in other categories

Table 2: Activity Dataset Subdomains and Descriptions

Label	Includes Discussion Around
Heart Health	heart rate, cardiovascular health, or related exercises
Lower Body Exercise	exercises targeting the legs, glutes, or lower body
Abdominal Exercise	exercises targeting the abdominal muscles (e.g., core work)
Upper Body Exercise	exercises targeting the arms, chest, shoulders, or upper body
Soreness	muscle soreness, recovery strategies, or pain management
Breathing	breathing techniques, exercises, or respiratory health
Posture	posture correction, exercises, or posture-related issues
Stretch	stretching routines, types of stretches, or flexibility
Rest	rest days, active recovery, or importance of rest
Personal Trainer	working with a personal trainer, their advice, or training plans
Weather	impact of weather conditions on activity or training
Injury	injuries, recovery from injury, or prevention strategies
Swimming	swimming routines, swimming techniques, or swimming benefits
Hiking	hiking trails, hiking gear, or hiking experiences
Marathon	marathon training, marathon preparation, or marathon races
Cycling	cycling routines, cycling techniques, or cycling benefits
Combat Sport	martial arts, boxing, or other combat sports
Yoga	yoga practices, yoga styles, or yoga benefits
Kettlebell	kettlebell training, exercises, or kettlebell benefits
Tracker	using activity trackers or wearable devices to monitor progress
Miscellaneous	any exercise and activity-related topic not covered in other categories

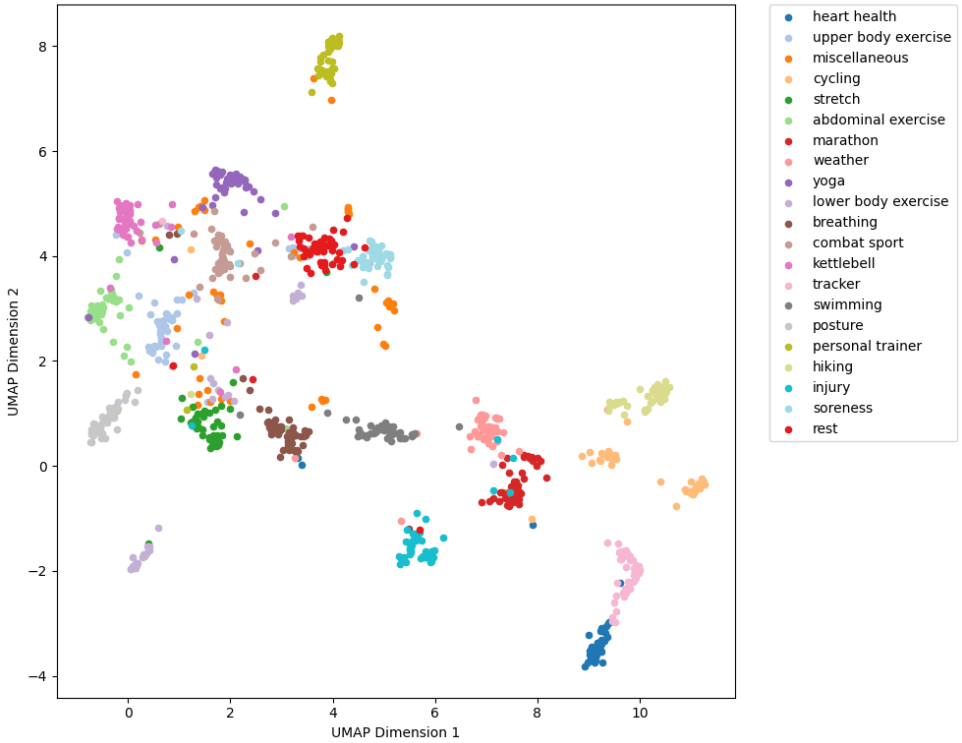

Figure 5: UMAP Visualization of BERT Embeddings for a Sample of Activity Subdomains (n=50 per class)

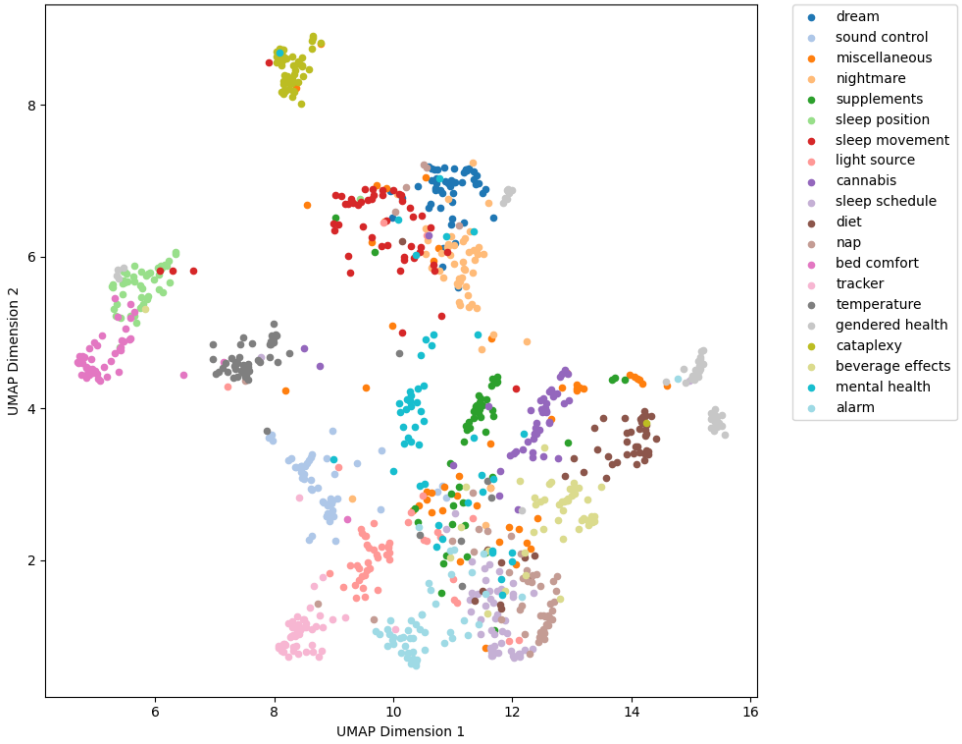

Figure 6: UMAP Visualization of BERT Embeddings for a Sample of Sleep Subdomains (n=50 per class)

Table 3: Classification Accuracy Across Tasks and Models (with error for Gemini models)

	Primary Classification									
Model	0	1	2	5	10	50	100	150	1000	All
BERT-base	-	0.43	0.59	0.68	0.81	0.94	0.97	0.96	0.98	0.99
DistilBERT	-	0.31	0.51	0.71	0.82	0.91	0.95	0.96	0.98	0.98
Electra-base	-	0.37	0.30	0.47	0.55	0.55	0.88	0.93	0.98	0.99
Electra-small	-	0.32	0.34	0.44	0.43	0.54	0.64	0.88	0.94	0.97
Gemini-nano	0.33 ± 0.00	0.30 ± 0.12	0.25 ± 0.02	0.29 ± 0.07	0.56 ± 0.02	-	-	-	-	-
Gemini-pro	0.87 ± 0.00	0.87 ± 0.01	0.84 ± 0.02	0.82 ± 0.03	0.84 ± 0.02	-	-	-	-	-
	Sleep Subdomain Classification									
Model	0	1	2	5	10	25	35	50	All	
BERT-base	-	0.06	0.08	0.30	0.60	0.79	0.84	0.88	0.87	
DistilBERT	-	0.05	0.14	0.43	0.64	0.82	0.82	0.84	0.84	
Electra-base	-	0.05	0.06	0.05	0.33	0.76	0.85	0.87	0.89	
Electra-small	-	0.06	0.05	0.09	0.09	0.37	0.66	0.75	0.77	
Gemini-nano	0.22 ± 0.01	0.27 ± 0.02	0.24 ± 0.01	-	-	-	-	-	-	
Gemini-pro	0.78 ± 0.00	0.78 ± 0.01	0.77 ± 0.02	-	-	-	-	-	-	
	Activity Subdomain Classification									
Model	0	1	2	5	10	25	35	50	All	
BERT-base	-	0.09	0.15	0.42	0.80	0.90	0.92	0.94	0.93	
DistilBERT	-	0.06	0.25	0.51	0.76	0.90	0.92	0.93	0.93	
Electra-base	-	0.05	0.07	0.17	0.53	0.85	0.92	0.92	0.91	
Electra-small	-	0.05	0.04	0.05	0.05	0.49	0.80	0.89	0.93	
Gemini-nano	0.27 ± 0.00	0.38 ± 0.02	0.40 ± 0.01	-	-	-	-	-	-	
Gemini-pro	0.82 ± 0.00	0.84 ± 0.00	0.85 ± 0.01	-	-	-	-	-	-	

Table 4: Synthetic Classification Accuracy Across Tasks and Models (with error for Gemini models)

	Primary Classification					
Model	0	1	2	5	10	60
BERT-base	-	0.15	0.47	0.65	0.88	0.96
DistilBERT	-	0.20	0.49	0.47	0.85	0.93
Electra-base	-	0.45	0.37	0.50	0.58	0.55
Electra-small	-	0.48	0.45	0.14	0.35	0.29
Gemini-nano	0.33 ± 0.00	0.42 ± 0.03	0.39 ± 0.02	0.40 ± 0.04	0.53 ± 0.13	-
Gemini-pro	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	0.97 ± 0.00	-
	Sleep Subdomain Classification					
Model	0	1	2	5	10	20
BERT-base	-	0.04	0.05	0.25	0.58	0.86
DistilBERT	-	0.02	0.06	0.38	0.60	0.90
Electra-base	-	0.02	0.06	0.09	0.30	0.85
Electra-small	-	0.05	0.05	0.05	0.10	0.32
Gemini-nano	0.24 ± 0.00	0.42 ± 0.03	0.44 ± 0.03	-	-	-
Gemini-pro	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	-	-	-
	Activity Subdomain Classification					
Model	0	1	2	5	10	20
BERT-base	-	0.07	0.08	0.18	0.55	0.89
DistilBERT	-	0.06	0.04	0.35	0.71	0.84
Electra-base	-	0.06	0.08	0.05	0.20	0.82
Electra-small	-	0.05	0.05	0.05	0.03	0.32
Gemini-nano	0.55 ± 0.00	0.67 ± 0.04	0.66 ± 0.01	-	-	-
Gemini-pro	0.97 ± 0.00	0.98 ± 0.01	0.98 ± 0.01	-	-	-

Table 5: Prompt Templates for Text Classification with Gemini. Refer to Table 1 and Table 2 for respective labels and descriptions.

Task	Prompt Template

Primary Classification (Zero-shot)	`You are a topic classifier. Your job is to classify the input post into one of 3 categories:`
	`- sleep`
	`- activity`
	`- out-of-distribution`
	`You must return only 1 of these 3 possible choices in one word.`
	`If the post is mostly about activity`
	`(e.g., [...list of activity labels...]), return 'activity'.`
	`If the post mostly pertains to sleep topics`
	`(e.g., [...list of sleep labels...]), return 'sleep'.`
	`If the post is mostly about neither, return 'misc'.`
	`Here is my post:`
	`Input: {post}`
	`Output:`

Continued on next page

Table 5 – continued from previous page

Task	Prompt Template

Primary Classification (Few-shot)	`You are a topic classifier. Your job is to classify the input post into one of 3 categories:`
	`- sleep`
	`- activity`
	`- out-of-distribution`
	`You must return only 1 of these 3 possible choices in one word.`
	`If the post is mostly about activity (e.g., [...list of activity labels...]), return 'activity'.`
	`If the post mostly pertains to sleep topics (e.g., [...list of sleep labels...]), return 'sleep'.`
	`If the post is mostly about neither, return 'misc'.`
	`Here are some examples:`
	`{''.join([f'Input: {examples[i]} Output: {labels[i]} ' for i in range(len(examples))])}`
	`Here is my post:`
	`Input: {post}`
	`Output:`

Sleep Sub-classification (Zero/Few-shot)	`You are a topic classifier. Your job is to classify the input post into one of the following sleep categories:`
	`{", ".join("{label}: {description}" for label, description in sleep_descriptions])}`
	`You must return only 1 of these possible choices.`
	`{Description for each sleep category}`
	`[Optional: Here are some examples: ...]`
	`Here is my post:`
	`Input: {post}`
	`Output:`

Activity Sub-classification (Zero/Few-shot)	`You are a topic classifier. Your job is to classify the input post into one of the following activity categories:`
	`{", ".join("{label}: {description}" for label, description in activity_descriptions])}`
	`You must return only 1 of these possible choices.`
	`[Optional: Here are some examples: ...]`
	`Here is my post:`
	`Input: {post}`
	`Output:`

Table 6: Pseudo-Code and Prompt Templates for Synthetic Dataset Generation Using Gemini Pro.

Task	Prompt Template

Sleep Posts	`for category, description in sleep_descriptions:`
	`  Generate a short online post of someone talking about {category}`
	`  as it pertains to sleep. The post should include discussion`
	`  around {description}.`

Activity Posts	`for category, description in activity_descriptions:`
	`  Generate a short online post of someone talking about {category}`
	`  as it pertains to activity and exercise. The post should include`
	`  discussion around {description}.`

Miscellaneous Sleep Posts	`Generate a short online post of someone talking about sleep.`
	`However, the post should NOT mention or have any content related to`
	`any of these topics:`
	`{", ".join(["{label}" for label in sleep_descriptions)}.`

Miscellaneous Activity Posts	`Generate a short online post of someone talking about activity and`
	`exercise. However, the post should NOT mention or have any content`
	`related to any of these topics:`
	`{", ".join(["{label}" for label in activity_descriptions)}.`

Out-of-Distribution Posts	`Generate a short online post of someone talking about anything but`
	`sleep, activity, and exercise topics. There should be no mention`
	`of personal fitness or wellness as it pertains to sleep or activity.`