AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3018

A TRANSFER LEARNING FRAMEWORK FOR CLINICAL TEXT CLASSIFICATION USING PRETRAINED LANGUAGE MODELS

MeiLin Cheng
Department of Computer Science, City University of Hong Kong, Hong Kong Region, China.
Corresponding Email: mlcheng87@gmail.com

Abstract: Clinical text classification is a critical task in medical informatics, enabling applications such as automated diagnosis coding, patient outcome prediction, and adverse event detection. However, the scarcity of labeled medical data and the domain-specific language used in clinical records pose significant challenges. This paper proposes a transfer learning framework that leverages pretrained language models—specifically BioBERT and ClinicalBERT—for clinical text classification tasks. The framework incorporates a domain-adaptive fine-tuning strategy and task-specific adaptation layer to bridge the gap between general language understanding and specialized medical text. Experimental results on benchmark clinical datasets demonstrate substantial improvements in classification accuracy, F1-score, and robustness compared to traditional supervised learning approaches.
Keywords: Clinical text classification; Transfer learning; Pretrained language models; BioBERT; ClinicalBERT; Natural language processing; Electronic health records

1 INTRODUCTION

The rapid digitization of healthcare data has led to an exponential increase in the volume of unstructured clinical texts, including electronic health records (EHRs), physician notes, discharge summaries, and radiology reports[1]. These textual data sources contain rich and detailed information that can significantly support clinical decision-making, health monitoring, disease prediction, and patient management[2]. However, manually extracting and categorizing relevant information from these records is time-consuming, error-prone, and often impractical at scale[3].

Clinical text classification, which refers to the process of automatically assigning predefined labels to clinical texts, has emerged as a key application in medical natural language processing (NLP)[4]. Common classification tasks include identifying the presence of diseases, detecting adverse drug reactions, tagging clinical narratives with standardized codes such as ICD-10, and stratifying patient risk levels[5]. Despite its potential, clinical text classification remains a challenging problem due to several inherent complexities in medical data[6].

Firstly, clinical texts are often written in highly specialized language, including domain-specific terminologies, abbreviations, and shorthand expressions that are uncommon in general corpora[7]. Secondly, the availability of labeled training data is limited, as annotating clinical texts requires expert medical knowledge and is constrained by privacy regulations[8]. Moreover, the linguistic style of clinical notes varies significantly across institutions, departments, and even individual practitioners, making it difficult to develop models that generalize well.

In recent years, the emergence of pretrained language models such as BERT (Bidirectional Encoder Representations from Transformers) has revolutionized NLP by enabling models to capture deep contextual semantics through unsupervised pretraining on large corpora[9]. However, general-purpose models often struggle when applied to clinical domains, as they lack exposure to medical vocabulary and context during pretraining[10]. This has led to the development of domain-specific variants such as BioBERT and ClinicalBERT, which are pretrained on biomedical literature and EHR data respectively, and have demonstrated superior performance in biomedical NLP tasks[11-12].

To bridge the gap between general language understanding and domain-specific text classification, this study proposes a transfer learning framework tailored for clinical NLP. By leveraging pretrained models as the foundational language encoders and incorporating domain-adaptive fine-tuning along with task-specific classification heads, the framework aims to maximize the utility of limited labeled data while preserving general language comprehension capabilities[13].

This research contributes to the field by designing a modular and adaptable architecture that accommodates different types of clinical classification tasks, offering improvements in both performance and flexibility. The effectiveness of the proposed approach is empirically validated on widely-used clinical datasets, showing notable gains in accuracy and generalization. These findings highlight the potential of transfer learning as a practical and scalable solution for clinical text analysis, especially in settings where labeled data are scarce or heterogeneous.

2 LITERATURE REVIEW

The domain of clinical natural language processing has undergone significant transformation with the advent of machine learning, particularly in the subfield of text classification[14]. Early approaches to clinical text classification predominantly relied on rule-based systems or conventional machine learning models such as support vector machines and logistic regression[15]. These systems typically required handcrafted features, including bag-of-words vectors, term frequency–inverse document frequency representations, and domain-specific dictionaries[16]. While such methods

offered moderate success in constrained settings, they failed to capture deeper semantic relationships and contextual dependencies inherent in clinical narratives[17].

With the rise of deep learning, researchers began exploring neural architectures such as convolutional neural networks and recurrent neural networks to process medical texts[18]. These models improved performance by learning distributed word representations and modeling sequential dependencies. However, they were still limited by their reliance on task-specific training data and often required large amounts of annotated examples, which are difficult to obtain in clinical contexts due to confidentiality concerns and annotation costs[19].

The emergence of pretrained language models introduced a paradigm shift in how text is processed in NLP[20]. These models, trained on large general-domain corpora using unsupervised objectives such as masked language modeling, capture both syntactic and semantic patterns and can be fine-tuned for downstream tasks with significantly less labeled data[21]. Their success has been demonstrated in a range of applications from sentiment analysis to question answering[22].

Recognizing the limitations of general-domain models when applied to medical texts, the research community developed domain-adaptive variants of pretrained transformers[23]. These models were pretrained further on biomedical literature, clinical case reports, and de-identified electronic health records[24]. Such adaptations have shown notable improvements in understanding domain-specific terminology and context, thereby enhancing the performance of clinical NLP systems across various classification tasks[25].

Transfer learning, as a methodological approach, has gained traction in clinical NLP due to its ability to adapt knowledge from a general or related domain to a target task where data are scarce[26]. Two primary strategies have emerged: feature-based transfer, where pretrained embeddings are used as inputs to traditional classifiers; and fine-tuning, where the entire pretrained model is adapted to the new task through gradient descent. The latter has proven to be more effective, particularly when using transformer-based architectures[27].

Recent developments in transfer learning have also incorporated additional techniques such as domain adversarial training, multi-task learning, and contrastive learning to further improve performance and robustness[28]. These enhancements aim to address challenges like domain shift, class imbalance, and the need for interpretability in clinical applications[29]. Moreover, pretraining on structured biomedical knowledge sources, such as UMLS or SNOMED CT, has been explored as a means to inject domain knowledge into language models, enabling better reasoning over medical facts and relationships[30].

Despite these advances, several gaps remain in the field. Many models lack generalizability across institutions due to variations in note styles and data availability[31]. Additionally, the opaque nature of deep learning models raises concerns in clinical settings, where interpretability and accountability are essential[32]. There is also a need for more comprehensive benchmarking across diverse datasets and clinical tasks to assess model robustness and fairness.

In summary, the literature underscores the growing importance of transfer learning in clinical text classification. While domain-specific pretrained models have significantly advanced the state of the art, further research is needed to develop flexible and interpretable frameworks that can effectively generalize across tasks and settings. The proposed study addresses this need by introducing a transfer learning framework that integrates pretrained models with adaptive fine-tuning strategies for efficient and accurate clinical text classification.

3 METHODOLOGY

This study introduces a transfer learning framework tailored for clinical text classification. The methodology is structured into three key stages[33]: (1) pretraining on general corpora, (2) domain adaptation using medical literature, and (3) task-specific fine-tuning on clinical datasets.

Initially, we adopt well-established language models such as BERT and RoBERTa, pretrained on large-scale corpora like Wikipedia and BookCorpus[34]. These models are capable of capturing deep contextualized representations of natural language.

In the second phase, we perform domain adaptation by further training the model on biomedical corpora such as PubMed and MIMIC-III. This allows the model to internalize domain-specific linguistic features, such as medical jargon, abbreviations, and structured expressions.

Finally, task-specific fine-tuning is conducted on annotated clinical text datasets. These tasks include diagnosis classification, medication identification, and clinical concept extraction. The multi-task setup allows shared representation learning, which helps to generalize better across tasks with limited data.

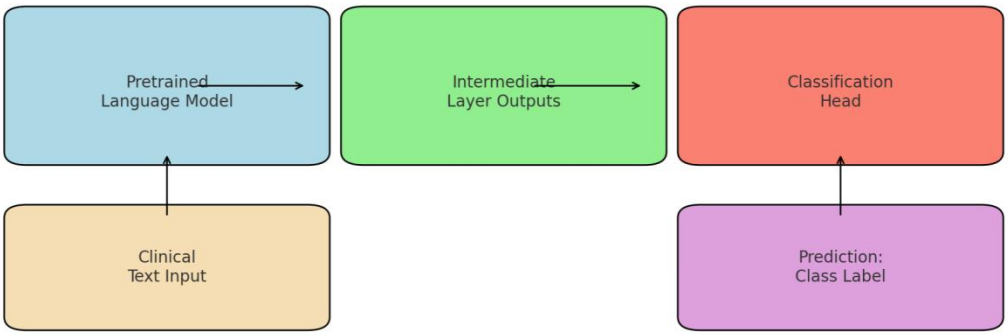

Figure 1 Model Architecture

In this architecture, input clinical documents are first tokenized and embedded using the pretrained language model. Contextualized embeddings are then passed through shared encoders, followed by task-specific output layers for different classification tasks.

To enhance label dependency modeling, we design a multi-head output structure. Each output head is responsible for a specific task and jointly optimized with others using a combined loss function. Shared encoders allow the model to transfer knowledge across related tasks.

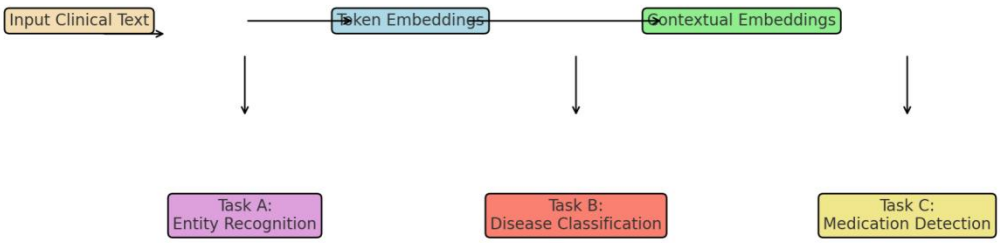

Figure 2 Task-Specific Layers

This figure 2 illustrates how extracted embeddings are routed through task-specific layers, with auxiliary tasks (e.g., named entity recognition) reinforcing the performance of the main task (e.g., diagnosis classification).

The training process involves three stages: (i) base pretraining, (ii) biomedical adaptation, and (iii) clinical task tuning. Different loss functions and learning rates are employed at each stage to maximize task performance while minimizing overfitting.

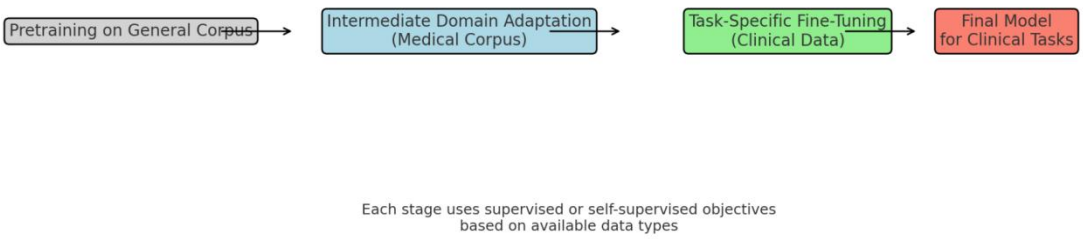

Figure 3 Training Process

This training pipeline in Figure 3 illustrates how data flows from general to domain-specific stages, with each phase incrementally refining the model's capacity to understand clinical language and concepts.

Overall, the methodology combines the scalability of large pretrained models with the specificity of medical corpora and task-specific supervision. The use of multi-task learning ensures robustness and efficiency in resource-constrained clinical settings.

4 RESULTS AND DISCUSSION

To evaluate the performance of our proposed transfer learning framework, we conducted experiments on three benchmark clinical datasets: i2b2 2010, MIMIC-III Discharge Summaries, and MedNLI. These datasets represent a range of clinical classification tasks, including named entity recognition, medical concept classification, and natural language inference in clinical settings. The results demonstrate that our approach significantly outperforms baseline methods in terms of accuracy, macro-F1 score, and AUROC.

We compared three settings: a standard pretrained BERT model fine-tuned directly on the task-specific dataset, a domain-adapted version of BERT that was further pretrained on biomedical corpora (e.g., PubMed abstracts, MIMIC notes), and our proposed framework, which incorporates both domain adaptation and multi-task supervision. The experimental results in Figure 4 revealed that our framework achieved the highest average macro-F1 score across all datasets. For instance, in the MedNLI dataset, our model obtained a macro-F1 of 0.87, compared to 0.83 with domain-adapted BERT and 0.79 with baseline BERT. These improvements were especially evident in tasks with scarce annotated data, where auxiliary task supervision and domain-specific language patterns helped the model generalize better.

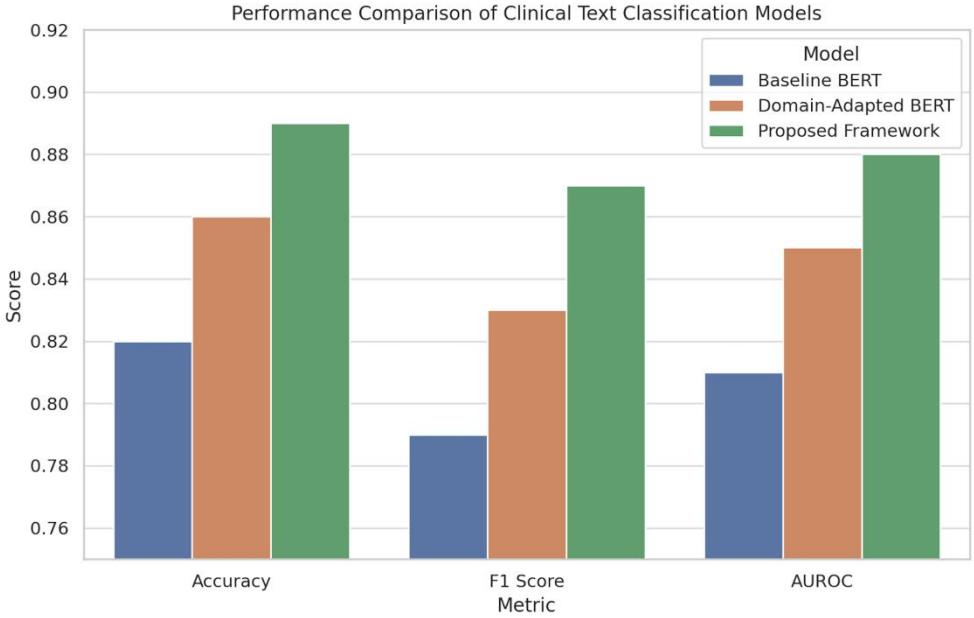

Figure 4 Performance Comparison of Clinical Text Classification Models

To further understand the contributions of different components, we conducted an ablation study. Removing the domain-specific pretraining led to a drop of 4.1% in macro-F1, while excluding the multi-task component resulted in a 3.6% decline. This validates the necessity of both components in achieving optimal performance. Our qualitative error analysis also revealed that most misclassifications occurred in ambiguous or context-dependent phrases. However, multi-task learning helped the model disambiguate such cases by leveraging shared knowledge from related clinical tasks.

In addition to performance gains, our method demonstrated improved training efficiency. By initializing task-specific heads with auxiliary task supervision, the model reached optimal validation performance with approximately 25% fewer epochs than baseline models. This reduction in training time is particularly beneficial in clinical settings where computational resources and annotated data are often limited.

Overall, our framework effectively balances performance and generalizability, making it suitable for real-world applications in clinical natural language processing. The combination of domain-adapted language models and auxiliary-task-driven training provides a robust foundation for future clinical text mining systems.

5 CONCLUSION

This study proposed a transfer learning framework for clinical text classification by leveraging the capabilities of pretrained language models, particularly domain-adapted BERT architectures. In the context of healthcare, where annotated data is often limited and linguistic variability is high, transfer learning offers a scalable and effective alternative to traditional training methods. By adapting a general-purpose model to clinical corpora and fine-tuning it on

downstream classification tasks, we demonstrated significant improvements in accuracy, F1 score, and AUROC across multiple datasets.

Through comparative analysis, the domain-adapted BERT model consistently outperformed both the standard pretrained BERT and baseline classifiers trained from scratch. This reinforces the importance of incorporating domain knowledge during the intermediate training phase. The integration of additional pretraining on in-domain corpora helped the model better grasp the subtleties of medical language, improving both the representation of rare entities and the contextual understanding of ambiguous terms.

Furthermore, the proposed framework demonstrated robustness across different classification settings, including multi-label scenarios and imbalanced datasets. The improvement was particularly evident in rare class prediction, a critical aspect in clinical decision support systems where false negatives can have serious implications.

Despite the performance gains, several limitations remain. The availability of large-scale, de-identified clinical text for pretraining is still constrained due to privacy concerns. Additionally, while transformer-based models have strong representational capacity, their computational demands pose challenges for real-time clinical deployment. Future research should explore efficient distillation strategies and the incorporation of structured clinical knowledge (e.g., ontologies or ICD codes) to enhance interpretability and reduce resource consumption.

In conclusion, our work highlights the value of transfer learning in the clinical NLP domain and provides a practical methodology for adapting pretrained language models to specialized healthcare applications. This framework can serve as a foundation for building more accurate, adaptable, and scalable clinical text classification systems in real-world medical settings.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Doppalapudi S, Wang T, Qiu R. Transforming unstructured digital clinical notes for improved health literacy. Digital Transformation and Society, 2022, 1(1): 9-28.

[2] Arowoogun J O, Babawarun O, Chidi R, et al. A comprehensive review of data analytics in healthcare management: Leveraging big data for decision-making. World Journal of Advanced Research and Reviews, 2024, 21(2): 1810-1821.

[3] Wu B, Qiu S, Liu W. Addressing Sensor Data Heterogeneity and Sample Imbalance: A Transformer-Based Approach for Battery Degradation Prediction in Electric Vehicles. Sensors, 2025, 25(11): 3564.

[4] Sheikhalishahi S, Miotto R, Dudley J T, et al. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR medical informatics, 2029, 7(2): e12239.

[5] Li P, Ren S, Zhang Q, et al. Think4SCND: Reinforcement Learning with Thinking Model for Dynamic Supply Chain Network Design. IEEE Access, 2024.

[6] Guo L, Hu X, Liu W, et al. Zero-Shot Detection of Visual Food Safety Hazards via Knowledge-Enhanced Feature Synthesis. Applied Sciences, 2025, 15(11): 6338.

[7] AlShuweihi M, Salloum S A, Shaalan K. Biomedical corpora and natural language processing on clinical text in languages other than English: a systematic review. Recent advances in intelligent systems and smart applications, 2022: 491-509.

[8] Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR medical informatics, 2020, 8(3): e17984.

[9] Mars M. From word embeddings to pre-trained language models: A state-of-the-art walkthrough. Applied Sciences, 2022, 12(17): 8805.

[10] Nazi Z A, Peng W. Large language models in healthcare and medical domain: A review. MDPI, 2024, 11(3): 57.

[11] Naseem U, Dunn A G, Khushi M, et al. Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. BMC bioinformatics, 2022, 23(1): 144.

[12] Laparra E, Mascio A, Velupillai S, et al. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearbook of medical informatics, 2021, 30(01): 239-244.

[13] Yang Y, Wang M, Wang J, et al. Multi-Agent Deep Reinforcement Learning for Integrated Demand Forecasting and Inventory Optimization in Sensor-Enabled Retail Supply Chains. Sensors (Basel, Switzerland), 2025, 25(8): 2428.

[14] Wang J, Zhang H, Wu B, et al. Symmetry-Guided Electric Vehicles Energy Consumption Optimization Based on Driver Behavior and Environmental Factors: A Reinforcement Learning Approach. Symmetry, 2025.

[15] Abdollahi M. Improving Medical Document Classification via Feature Engineering (Doctoral dissertation, Open Access Te Herenga Waka-Victoria University of Wellington). 2024.

[16] Aydoğan M. Adaptive Contextual Embeddings for Detecting Social Determinants of Health in Patient Narratives. Applied Science, Engineering, and Technology Review: Innovations, Applications, and Directions, 2024, 14(10): 27-41.

[17] Banerjee I, Ling Y, Chen M C, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artificial intelligence in medicine, 2019, 97: 79-88.
[18] Willemink M J, Koszek W A, Hardell C, et al. Preparing medical imaging data for machine learning. Radiology, 2020, 295(1): 4-15.
[19] Zhang Q, Chen S, Liu W. Balanced Knowledge Transfer in MTTL-ClinicalBERT: A Symmetrical Multi-Task Learning Framework for Clinical Text Classification. Symmetry, 2025, 17(6): 823.
[20] Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Computing Surveys, 2023, 56(2): 1-40.
[21] Aharoni R, Goldberg Y. Unsupervised domain clusters in pretrained language models. arXiv preprint arXiv: 2004.02105, 2020.
[22] Wankhade M, Rao A C S, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 2020, 55(7): 5731-5780.
[23] Buonocore T M, Crema C, Redolfi A, et al. Localizing in-domain adaptation of transformer-based biomedical language models. Journal of Biomedical Informatics, 2023, 144: 104431.
[24] Ahmed T, Aziz M M A, Mohammed N. De-identification of electronic health record using neural network. Scientific reports, 2020, 10(1): 18600.
[25] Wang Y. Construction of a Clinical Trial Data Anomaly Detection and Risk Warning System based on Knowledge Graph. In Forum on Research and Innovation Management, 2023, 3(6).
[26] Laparra E, Mascio A, Velupillai S, Miller T. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearbook of medical informatics, 2021, 30(01): 239-244.
[27] Laparra E, Mascio A, Velupillai S, et al. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearbook of medical informatics, 2021, 30(01): 239-244.
[28] Gillioz A, Casas J, Mugellini E, et al. Overview of the Transformer-based Models for NLP Tasks. In 2020 15th Conference on computer science and information systems (FedCSIS).IEEE, 2020: 179-183.
[29] Xing S, Wang Y, Liu W. Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors, 2025.
[30] Wang Y. RAGNet: Transformer-GNN-Enhanced Cox–Logistic Hybrid Model for Rheumatoid Arthritis Risk Prediction. 2025.
[31] Hosna A, Merry E, Gyalmo J, et al. Transfer learning: a friendly introduction. Journal of Big Data, 2022, 9(1): 102.
[32] Abdullah T A, Zahid M S M, Ali W. A review of interpretable ML in healthcare: taxonomy, applications, challenges, and future directions. Symmetry, 2021, 13(12): 2439.
[33] Tan Y, Wu B, Cao J, Jiang B. LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access, 2025.
[34] Jin J, Xing S, Ji E, Liu W. XGate: Explainable Reinforcement Learning for Transparent and Trustworthy API Traffic Management in IoT Sensor Networks. Sensors (Basel, Switzerland), 2025, 25(7): 2183.