Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–21

DOI: 10.1177/20552076241287357
journals.sagepub.com/home/dhj

Sage

Transformer-based active learning for multi-class text annotation and classification

Muhammad Afzal[1,†], Jamil Hussain[2,†], Asim Abbas[3,4], Maqbool Hussain[5], Muhammad Attique[6,*] and Sungyoung Lee[7]

Abstract

Objective: Data-driven methodologies in healthcare necessitate labeled data for effective decision-making. However, medical data, particularly in unstructured formats, such as clinical notes, often lack explicit labels, making manual annotation challenging and tedious.

Methods: This paper introduces a novel deep active learning framework designed to facilitate the annotation process for multiclass text classification, specifically using the SOAP (subjective, objective, assessment, plan) framework, a widely recognized medical protocol. Our methodology leverages transformer-based deep learning techniques to automatically annotate clinical notes, significantly easing the manual labor involved and enhancing classification performance. Transformer-based deep learning models, with their ability to capture complex patterns in large datasets, represent a cutting-edge approach for advancing natural language processing tasks.

Results: We validate our approach through experiments on a diverse set of clinical notes from publicly available datasets, comprising over 426 documents. Our model demonstrates superior classification accuracy, with an F1 score improvement of 4.8% over existing methods but also provides a practical tool for healthcare professionals, potentially improving clinical documentation practices and patient care.

Conclusions: The research underscores the synergy between active learning and advanced deep learning, paving the way for future exploration of automatic text annotation and its implications for clinical informatics. Future studies will aim to integrate multimodal data and large language models to enhance the richness and accuracy of clinical text analysis, opening new pathways for comprehensive healthcare insights.

Keywords

Text classification, text annotation, active learning, transfer learning, deep learning, BERT, clinical text, SOAP

Submission date: 13 July 2023; Acceptance date: 10 September 2024

Introduction

In today's world, patient data are logged into an electronic health record (EHR) system in both structured and unstructured formats.[1] The unstructured form mainly includes clinical notes, discharge summaries, and diagnostic test reports written in natural language. These reports contain vital information that might help solve clinical questions about patient health conditions, clinical reasoning, and inferencing. However, due to the time limitation, physicians have difficulty examining the unstructured information at the point of care.[2] Traditionally, clinically relevant information from clinical documents is extracted through manual methods with the support of clinical domain experts,

[1]College of Computing, Birmingham City University, Birmingham, UK
[2]Department of AI and Data Science, Sejong University, Seoul, Korea
[3]Department of Computer Science, St John's University, Jamaica, NY, USA
[4]School of Computer Science, University of Birmingham, Birmingham, UK
[5]School of Computing and Engineering, University of Derby, Derby, UK
[6]Department of Software, Sejong University, Seoul, Korea
[7]Department of Computer Science and Engineering, Kyung Hee University, Yongin, Korea

[*]Current affiliation: Department of Artificial Intelligence, Ajou University, Suwon-Si, South Korea.
[†]These authors contributed equally to this work.

Corresponding author:
Sungyoung Lee, Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, Korea.
Email: sylee@oslab.khu.ac.kr

which creates hurdles in terms of scalability and costs. At the same time, data availability allows researchers to execute automated algorithms extracting helpful information for efficient disease care.[3] (NLP) plays a significant role in the clinical domain for various applications, such as medical concept identification in different clinical documents.[4] Recently, NLP applications have further diversified to use for disease outbreak detection, conversion of free text to structured features for decision support, answering clinical questions, and accessing knowledge embodied in free-text clinical and biomedical resources.[5]

The information extraction facilitated with NLP led to automated clinical text classification in clinical predictive analytics that emerged with the huge creation of clinical notes and speedily growing adoption of EHR systems.[6] Two types of techniques: symbolic and statistical machine learning, are commonly used for clinical text classification tasks.[7] Symbolic techniques are used in applications that involve hand-crafted rules by domain experts, like logic rules and regular expressions. Although rule-based methods are effective in the clinical domain because of sub-language properties, it can be laborious to develop a system that requires collaboration between technical NLP experts and clinical domain experts. Moreover, the final applications may have limitations of portability and generalization beyond the scenario for which it was intended.[8]

Machine learning (ML) methods have been proven to be efficient for the tasks of clinical text classification. However, an effective supervised ML model still needs human involvement to annotate a huge set of training data. The efforts by domain experts to unstructured label data are a significant blockade of inefficient data analysis.[9] The annotation problem is of primary focus in the medical domain because of the lack of clinical data available to the public and expert knowledge for accurate annotations. The other popular methods, such as crowdsourcing, are unsuitable for creating labeled clinical training data because of the sensitive nature of the domain. Also, the findings of a systematic review[9] show that most datasets used in training ML models for text classification consist of mere hundreds or thousands of records because of annotation blockade.

The manual annotation process issues have been tried to be resolved by modern orthogonal approaches such as active learning (AL) and transfer learning (TL), which are utilized as machine-assisted pre-annotation methods.[10] AL provides a subset of high-value training samples by reducing the huge data required for labor-intensive data annotation without losing the quality.[11] The initial data for the AL process can be prepare through symbolic techniques, such as a rule-based approach combined with a domain- or task-specific lexicon or dictionary like UMLS[12] and Bioportal.[13] The selection of samples is iterative starting with a high-quality manually annotated subset of samples and moving to automatically generate another subset of annotations, thus increasing the subset-to-annotated text for use in the subsequent iterations of the process.[10] AL approaches have been applied in a clinical domain to decrease labor-intensive data annotation burden and enhance the model classification performance with a few labeled examples sets.[11–14]

In recent times, we have seen a growing amount of biomedical data available in textual form. Substantial advances in the development of pretraining language representation models provide an opportunity for a range of biomedical domain tasks, such as pretrained word embedding, sentence embedding, and contextual representations. According to Beltagy et al.,[15] the SciBERT outperforms the baseline encoder representations from transformers (BERT) model on biomedical tasks. SciBERT is a deep learning-based language model that uses the original BERT model and is trained on scientific articles for the biomedical domain.

Given the inherent difficulties in clinical text annotation and classification, this work employs a mixed-method design that combines experimental and observational research components. Our methodology starts by creating a rule-based system to produce a seed dataset, which is subsequently utilized to initialize an AL model based on transformer mechanisms. Implementing this iterative procedure not only reduces the amount of annotation required but also improves the learning efficiency of the model. Through the utilization of AL and TL approaches, our methodology deliberately chooses and defines data points that optimize the performance of the model. These results indicate a substantial enhancement in the accuracy of categorizing clinical notes. This study proposes a methodology for clinical text annotation and classification by combining AL and TL learning approaches to minimize human efforts in creating labeled data. The primary challenges in supervised ML involve data annotation and how AL can alleviate these obstacles. Specifically, manually annotating data poses a significant challenge due to its time-consuming and labor-intensive nature, often leading to bottlenecks in large-scale NLP projects. In contrast, AL offers a strategic solution to this problem by selectively querying unlabeled data that, once annotated, greatly benefits the model's learning process. This approach not only streamlines the annotation process but also boosts the model's performance with a potentially smaller, yet richer, dataset. Tackling these challenges directly enables a comprehensive understanding of the trade-off between the labor costs of annotation and the efficiency improvements provided by AL, establishing a solid foundation for further investigation of our research contributions.

The proposed methodology employed a rule-based NLP algorithm based on a lexical approach that automatically annotates the unlabeled input data to create an initial seed dataset. Using the initially labeled dataset, we design an AL approach by training transformer-based deep learning

to enhance the initial seed data. The AL output, that is, the enhanced annotated data are used to train the proposed SciBERT-based multiclass classification model to classify texts in the clinical documents into four classes of the SOAP (subject, object, assessment, plan) protocol.[5] SOAP is a well-known structure used for patient information organized into four logical compartments.

To demonstrate the usefulness of the proposed methodology, we conducted a set of experiments on clinical notes acquired from a public dataset (i2b2/VA 2010).[16] The findings of the proposed approach indicate a significant reduction in annotation costs by achieving higher accuracy compared to the existing approaches used for the same task in the past. Furthermore, our approach is unique by applying novel AL methodology enhanced with TL for embedding to perform text classification tasks using an attention-based deep learning model. This approach is different from traditional NLP approaches in terms of context capturing within SOAP sections. For instance, a medication "xyz" may appear in a clinical note in two different forms; "xyz" is used currently, and "xyz" is prescribed for future use. Here identifying medication names correctly is not sufficient, but the context is important too. Identifying SOAP sections differentiates between the "xyz" medication as currently in use (subjective) and prescribed for the future (plan).

Our proposed approach provides an end-to-end solution involving clinical text preprocessing, a rule-based model for initial data annotations, a deep AL-based model for enhanced data annotations, and multiclass classification model development, validation, and testing. The proposed methods are not only useful for clinical text classification but other NLP tasks and applications, such as question-answering systems, clinical decision support systems, clinical follow-up systems, and health technology assessment processes. Generally, the automatic clinical annotations and labeling created with our proposed models are helpful for any clinical text classification or prediction task that needs labeled data. In summary, the key contributions of this study are as follows:

- Developing syntactic and semantic algorithms for unstructured clinical text preprocessing and section identification to prepare initial training data with SOAP labels as seed data for the AL model
- Developing a robust transformer-based AL model with uncertainty-based sampling—least confidence query strategy—for annotating unlabeled clinical data with SOAP labels
- Developing a dual attention network model, which employs two inputs: (a) SciBERT-based transfer learning (TL) input for capturing contextual information and (b) UMLS-based semantic enrichment (UMLS-SE) input to help capture semantic information

Literature review

There are two types of experimental settings available in the clinical domain: shared-task settings and clinical practice settings. In shared task settings, challenging NLP[16] corpora are typically made accessible with well-defined evaluation methods and public availability; hence, they are commonly recognized as benchmarks. However, in a clinical practice setting, the EHR is used directly for idea extraction in real-world contexts, such as internal medicine and orthopedics.

In a shared-task setting, it is more difficult, expensive, and time-consuming to construct a ML-based concept extraction application as there is insufficient annotated data. In a clinical practice context, clinical information extraction and classification tasks are performed using symbolic and statistical ML, as stated in the introduction. The current AL technique has resolved the problem of automated data annotation. To employ AL strategies, initial data is created using a symbolic method and domain-specific terminology. In the study, MedCATTrainer,[11] a web-based interface to extract medical concepts from EHR free text is developed. They obtained the initial semantic annotation from UMLS,[12] an open-source biomedical ontology repository, as well as rule patterns for concept identification, and afterward stored the annotated data in the database.

The interface enables a user to semantically edit annotated concepts or contribute semantic annotation to a missing concept, which they refer to as the AL technique. After getting many annotated ideas, an ML model such as a random forest is employed.

However, the early findings have not been presented by the author in the paper, and domain expertise is required to effectively run this application and annotate medical concepts. Word embedding similarity is another technique that plays a key part in the AL process. A model that has been pretrained is used to create the embedding of labeled and unlabeled data. The embedding similarity between labeled and unlabeled data is then assessed. Within a certain embedding similarity threshold value, unlabeled data are classified into a label data category.

In their research, Hussain et al.[17] have suggested a unique approach for identifying causal relationships in clinical text. Initial data are created using a symbolic approach, and a Google News word2vec pretrained[18] model is used for semantic expansion. Using BERT, the extended causal terms are turned into an embedded vector afterward. These embedded vectors are then used to calculate a cosine similarity matching score against causal words contained in two additional datasets. Finally, the domain expert verifies the predicted words from different datasets, concluding the AL process.

Ning An et al.[19] used word embedding with cosine similarity to detect causal relationships as a four-class classification problem. One-hot encoding converts causal verbs in the

seed list and verbs in NP-VP-NP ternaries into encoding vectors. Based on a Wikipedia dataset, these vectors are translated using continuous Skip-Gram. The encoded vectors are compared using cosine similarity, and the pair with the most similarity over 0.5 is used to classify the causal relationship and update the seed list. This technique earned an F-score of 78.67%, a substantial improvement over earlier causal link detection efforts.

Li et al.[20] have used AL to reduce annotation requirements in the deidentification workflow by incorporating real clinical trials and i2b2 datasets to show e improved performance of trained models compared to the traditional passive learning framework.

Similarly, Tomanek and Hahn[21] examined the impact of AL in decreasing the time required for data annotation for entities (person, organization, and location) extraction. They noticed that the AL process significantly decreases up to 33% data annotation time and cost compared to baseline. Chen[22] conducted a simulation experiment to reannotate a subset of the i2b2/VA 2010 dataset from the concept extraction challenge. Their results showed that the AL-based query strategy reduced the volume of data needed for manual annotation compared to baseline.

AL is used in other domains such as sentiment analysis,[23] where the authors proposed a novel active deep network (ADC) to solve the problem of the small dataset in the sentiment classification problem. In another study by Hajmohammaadi et al.,[24] they used AL and self-training for cross-lingual sentiment classification and other baseline models to check the effectiveness of their proposed model; they found that AL performed better when compared with baseline models (without using AL).

In addition to AL, researchers have used TL to learn knowledge from previously learned domains and apply it to newer domains and tasks. Most real-world applications suffer from data deficiency that results in suboptimal models based on deep learning approaches. TL is touted to address this issue by allowing pretrained models from domain A to be applied to tasks in another domain B; both A and B are related domains. TL is the dominant approach leveraged by leading language models such as RNNs, LSTMs, and transformer-based language (TBL). These models can be used for any downstream task, language, or domain. The TBL models perform better on various NLP tasks as compared with other models. In modern NLP techniques, the researcher combines TL methods with large-scale TBL models to achieve better performance. The existing language models based on RNNs and LSTMs suffer the vanishing gradient problem and cannot handle the longer contextual dependencies.

The LSTM-based models, such as ELMO (embeddings from language model) or ULMFiT (universal language model fine-tuning) are still used for modern NLP tasks. Still, the main limitations of LSTM-based models are challenging to train in a parallel way.

The word representation of the ELMO contains richer information compared to a standard or traditional word embedding such as Skip-Gram[25] and global vectors for word representation (GloVe).[26] Although the ELMO model is shown to have a good performance in some name entity recognition (NER) tasks, such as the CoNLL 2003 NER task, it is trained in a general domain and, as a result, does not demonstrate the desired performance for a clinical concept extraction task.

The transformer architecture resolves these issues by an attention mechanism, which creates an entire sequence from the whole document and trains the model in a parallel fashion. Various TBL models with slight differences exist for modern NLP tasks, but the performance of BERT-based models is exceptional.[27]

BioBERT[28] and ClinicalBERT[29] are recent examples of domain adaptations of BERT. BioBERT is trained on PubMed abstracts and PMC full-text publications, while ClinicalBERT is trained on MIMIC-III clinical text.[29] SciBERT[15] is trained on the complete text of 1.14 million biomedical and computer science publications from the Semantic Scholar corpus to increase performance on subsequent scientific NLP tasks. The SciBERT is assessed for five fundamental NLP tasks, including NER, participants, interventions, comparisons, and outcomes (PICO) in a clinical trial publication,[30] text classification, relation classification, and dependency parsing. We used SciBERT for SOAP label classification, a clinical protocol used for patient information management into four logical compartments because we believe SciBERT has already been evaluated on PICO, a clinical protocol used for clinical questions in terms of problem, intervention, comparison, and outcome.

So, the literature review highlights the intersection of deep learning and NLP as a frontier of innovation in the clinical domain, demonstrating the potential for significant advancements in healthcare delivery and patient care. The evolution from traditional NLP techniques to the adoption of advanced methodologies like AL, TL, and the integration of domain-specific transformer-based models signifies a transformative shift towards more accurate, efficient, and nuanced processing of clinical data.

Methodology

This section describes the proposed framework of SOAP-based data labeling and classification of clinical text. The framework is divided into three steps, as shown in Figure 1. In the first step, a rule-based algorithm ("SOAPNotesParser") is employed for initial data labeling (seed data annotations). According to the SOAP protocol, the rule-based algorithm includes both syntactic and semantic approaches to annotate different sections in the clinical notes. In the second step, an AL model is designed to create more data with SOAP labels as a training dataset for the classification model. Finally, a pretrained model is used to create embeddings to enrich the training data for attaining

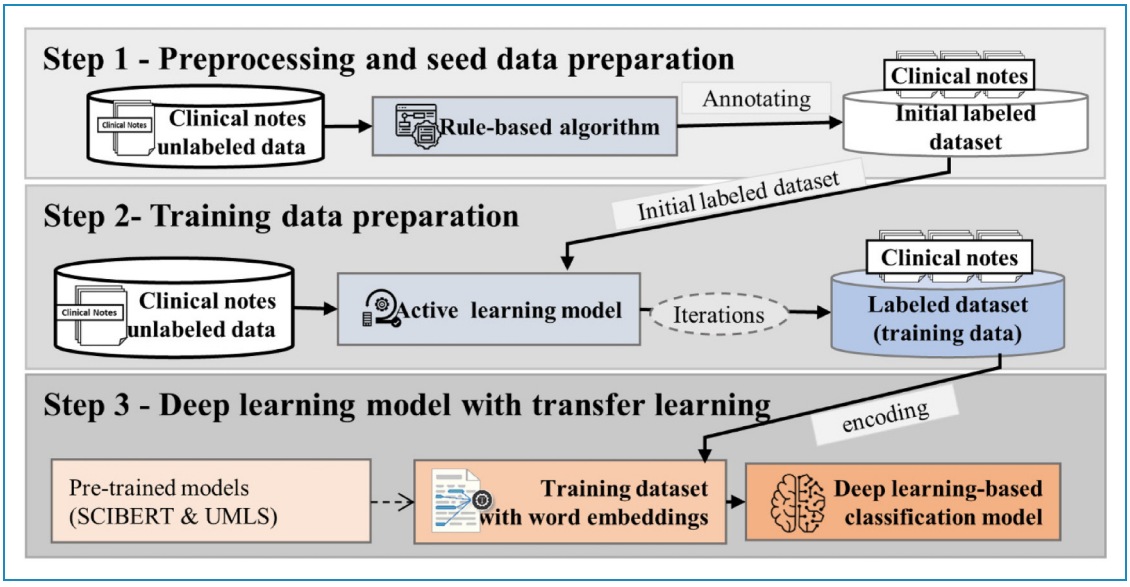

Figure 1. SOAP-based data labeling and classification framework of unstructured clinical notes.

Table 1. Dataset sources along with the number of clinical notes.

Dataset Source	Clinical Notes
Partners healthcare	97
Beth Israel deaconess medical center	73
i2b2 national center	256
Total	426

data and gaining maximum throughput out of the final deep learning model, which we eventually utilize to classify the unseen clinical notes.

Dataset

This study was conducted using a dataset composed of unstructured clinical discharge summaries collected from three key sources: the i2b2 National Center, Partners Healthcare, and Beth Israel Deaconess Medical Center, as shown in Table 1. The dataset's comprehensive breakdown is as follows, highlighting the number of clinical notes and the distribution of labeled versus unlabeled data. Partners Healthcare consists of 97 clinical notes, Beth Israel Deaconess Medical Center contains 73 clinical notes, and the dataset provided by the i2b2 National Center for System Evaluation contains 256 clinical notes. Cumulatively, we utilized 426 unstructured clinical discharge summaries in the proposed methodology. These clinical notes consist of explicitly defined sections used for section-based SOAP annotation.

The clinical notes encompass a variety of explicit and implicit sections, meticulously annotated to align with the SOAP framework. This approach ensures a structured analysis and classification of the clinical text.

Initial label dataset for the active learning process: Annotation and preparation

We developed and implemented an algorithm ("SOAPNotesParser") to efficiently parse and label clinical notes according to the SOAP framework for the AL process as shown in Figure 2. In this step, we selected 20 clinical notes having explicit header sections. This process is designed to transform unstructured clinical text into organized data, facilitating the AL process.

This initial dataset was intentionally diverse, spanning various document types, medical specialties, and patient demographics to ensure broad representation. Selected clinical notes were those with high annotation confidence by the "SOAPNotesParser" and input from domain experts, focusing on the inclusivity of both common and rare conditions and consideration for evolving medical practices.

The algorithm consists of the following steps.

1. *Identifying the SOAP sections.* The initial step involves scanning the clinical note for indicators of the main SOAP sections: subjective, objective, assessment, and plan. These sections are integral to the structure of clinical documentation, each serving a distinct role in encapsulating different aspects of patient care. By recognizing the keywords or phrases that typically denote the beginning of each section, the algorithm effectively demarcates the boundaries of these categories within the text.

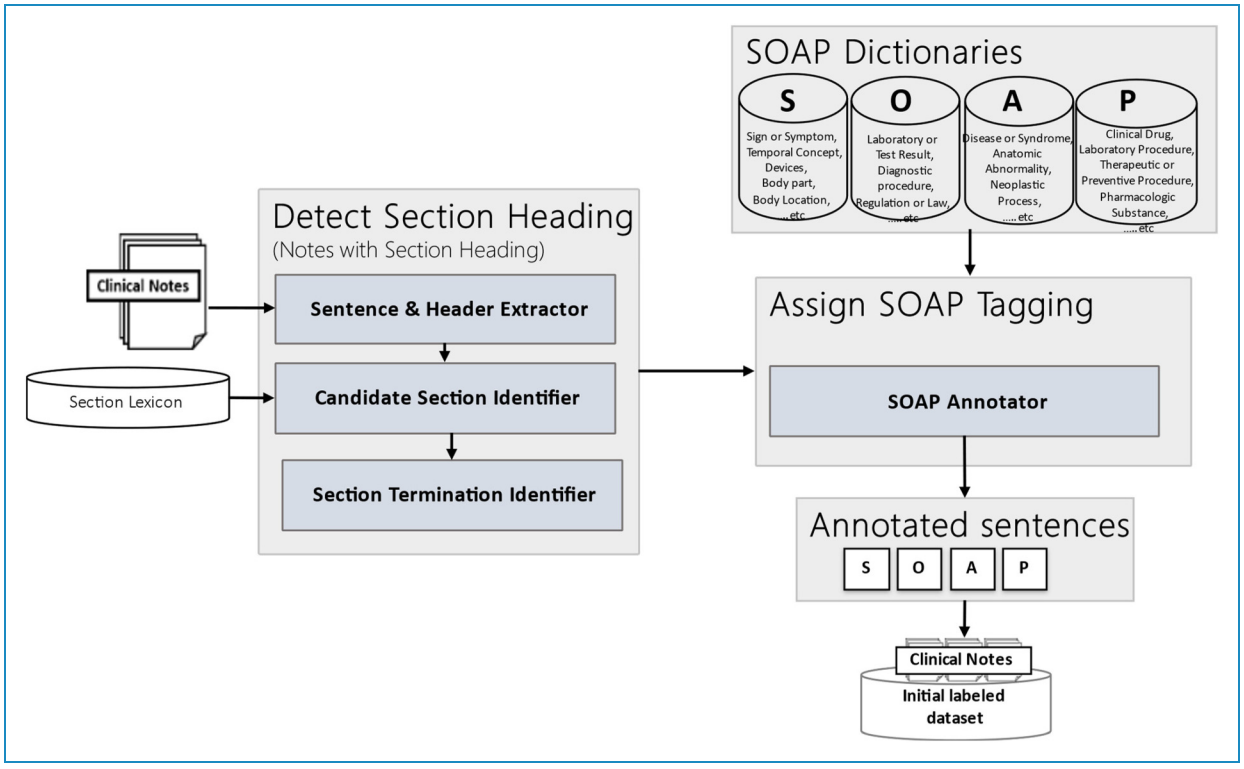

Figure 2. Workflow of SOAP section detection and annotation in clinical notes.

2. *Accumulating text under each section*. Once a section header is identified, the algorithm accumulates text corresponding to that section. It captures the content line by line, aggregating it until a new section header is encountered. This ensures that all information pertinent to a particular aspect of the SOAP framework is grouped, maintaining the integrity and context of the original clinical note. Importantly, the algorithm skips the header line itself to avoid redundancy, focusing instead on the substantive content that follows.
3. *Assigning labels to text*. As the algorithm aggregates text under each SOAP section, it also assigns appropriate labels to this content, indicating whether it pertains to subjective, objective, assessment, or plan aspects of patient care. This labeling is crucial for downstream applications, providing a clear, structured framework for analyzing the note's content. The process distinguishes between different types of clinical information, from patient-reported symptoms to treatment plans, enhancing the utility of the extracted data.
4. *Handling subheadings and complex structures*. The algorithm is adept at navigating the complexities of clinical documentation, including various subheadings and nuanced formatting that may occur within each main SOAP section. By employing a flexible parsing strategy, it can accommodate diverse document structures, ensuring comprehensive and accurate categorization of the text. This capability is particularly important given the wide range of documentation styles and conventions used across different healthcare settings. Some parts of this work can be referred from our previous work[31] and the section header terminology lexicon (SHTL) is based on works.[32]
5. *Producing structured output*. The culmination of the parsing and labeling process is the generation of structured output. The algorithm converts the categorized text into a format that is amenable to further analysis, such as a list of dictionaries. Each entry in the output indicates the section to which the text belongs, along with the labeled content itself.

Preprocessing and auto-labeling using the active learning process

The initial dataset from the 20 clinical notes obtained produced around 243 label instances as training datasets for the AL process. So, to label the remaining clinical notes (406 in total), initially, we employed preprocessing, where clinical notes were segmented into individual sentences. This segmentation was executed based on specific rules: a newline character ("\n") or the occurrence of a period followed by a space and an uppercase letter, indicative of the start of a new sentence. Following sentence segmentation,

two noteworthy observations were made: the prevalence of short sentences (those with fewer than a prespecified number [<5] of words) and duplicate sentences. These characteristics can be attributed to the uniformity in documenting physical and medical examinations, along with the concise way medical records are often completed by healthcare professionals.

To obtain a cleaner dataset, we filtered out short sentences and duplicates with preprocessing. The natural language toolkit (NLTK) again used it to convert the words by finding tokens out of them and excluded sentences that had less than five words (not very helpful information can be drawn from a sentence with as few as five or so words). Through the setting of a threshold on size, we ensured that all other sentences had the most information. It also removed duplicate sentences, so it reduced the text pattern for unique content. This was critical to improving the performance of AL in clinical note classification, as well-curated examples improved relevance and accuracy. For this study, we used an AL approach using a small-text framework to choose the most insightful unlabeled data from the pool.[33]

To further elaborate on the AL process, once the initial classifier is trained using the small seed initial dataset, it employs the small-text framework[33] for executing the pool-based sampling with the least confidence query strategy as shown in Figure 3. This technique involves presenting the model with unlabeled data and asking it to predict labels for these instances. Those with the lowest confidence in their predictions are deemed the most valuable for learning because they represent the boundary cases about which the model is most uncertain.

The selected instances are then reviewed by an oracle—a human expert or an automated system capable of providing the correct labels. This step is crucial as it ensures that the model is trained on accurately labeled data, thereby enhancing its learning efficiency. Once the oracle annotates the chosen instances with the correct labels, these newly labeled examples are added to the training dataset, and the model is retrained. This iterative cycle of prediction, selection by least confidence, and retraining with newly labeled data continues until the model reaches a state of convergence. Convergence is defined as the point at which additional training on new data does not significantly improve the model's performance, indicating that the model has achieved its maximum learning potential given the available data.

During this iterative process, the use of SciBERT,[15] a pretrained transformer-based model specifically tailored for scientific text, plays a pivotal role. The choice of SciBERT, with its uncased variant, allows for the construction of high-quality embedding vectors from the small-label dataset. These embeddings capture the semantic nuances of the scientific domain, enabling more effective model training than would be possible with general-purpose language models.

The AL methodology detailed in this study underscores the efficiency of using a targeted approach to data annotation. By focusing on instances where the model's certainty is lowest, the AL strategy ensures that the model's training is both efficient and effective, reducing the need for a vast amount of labeled data.[25] This approach is particularly beneficial in domains where labeled data is scarce or expensive to obtain, such as specialized scientific fields.

Moreover, the adoption of a pool-based sampling strategy, as opposed to stream- or membership-based selection, is motivated by the practical considerations of having a relatively small-label dataset and a substantially larger pool of unlabeled data.[34] The pool-based approach allows for a more systematic exploration of the data space, ensuring that the model encounters a diverse set of examples during its training. This diversity is critical for developing a robust model capable of generalizing new, unseen data well.

In conclusion, the AL approach described here leverages the strengths of the small-text framework, SciBERT embeddings, and a judicious selection strategy to efficiently tackle the challenge of text annotation in a data-scarce environment. The methodology's emphasis on targeting model uncertainty and iteratively refining the training dataset through expert annotation leads to a significantly improved model performance. This approach not only accelerates the process of model development but also enhances the model's accuracy and generalizability, making it a valuable

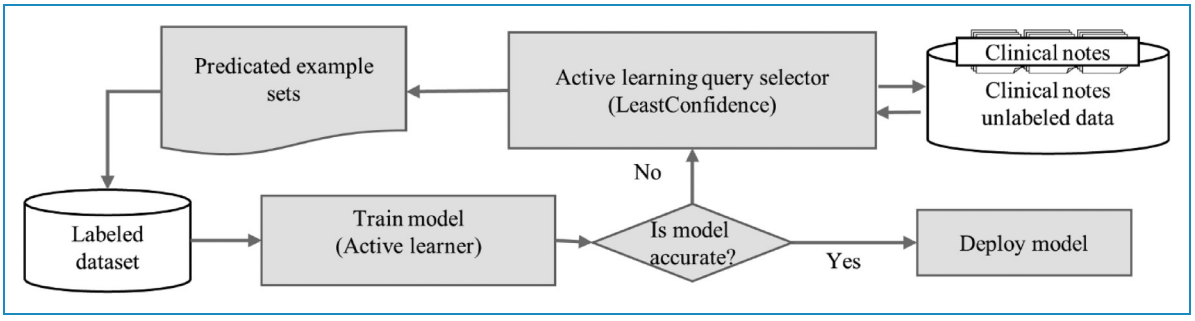

Figure 3. A step-by-step process of automatic text annotation using an active learning approach.

strategy for advancing ML applications in specialized domains. By the end of this process, we successfully labeled the 3146 instances as a training dataset.

SOAP-BioMedBERT—The proposed model

With the AL model, we add enough labeled records to the dataset, which is sufficient to use as a training dataset for a state-of-the-art deep learning model. Furthermore, we developed an attention-based deep learning model named "SOAP-BioMedBERT." A high-level workflow architecture of the proposed model for classifying clinical notes with SOAP labels is depicted in Figure 4.

The proposed model utilizes TL and UMLS-based semantic enrichment (UMLS-SE) to achieve optimal results. The combination of the two networks was intended to help capture both contextual and semantic information in clinical notes.

In the model, the weight-tuning operation is activated along with the SOAP-based training dataset to learn specific characteristics of the data. Firstly, the clinical text is normalized using data preprocessing techniques such as removing accented characters, expanding contractions, removing special characters, stemming, and removing stop words. Then, the normalized clinical text is inputted into two proposed networks for predicting the final SOAP label. Both networks combine concatenation, dropout, and dense layers using the SoftMax activation function. The cross-entropy loss is optimized using Adam and a dropout of 0.3.

Contextual information network. Our network is meticulously architected to capture the intricate context of clinical text, employing three distinct layers: word embedding, encoding, and attention layer. We utilize the pretrained SciBERT-based uncased model,[15] which operates on a BERT-based architecture with 24 layers, at the word embedding stage. This transformer-based model, initially representing words in their embedded form, employs multiheaded attention across each layer to iteratively refine word representations, informed by the surrounding textual context as shown in Figure 4(a).

The BERT architecture is adept at capturing bidirectional contextual cues. However, the clinical domain's nuanced linguistic structure demands enhanced processing capabilities. Hence, we enrich the SciBERT embeddings with a bi-directional long short-term memory (Bi-LSTM) network. This addition strategically augments the model's capacity to discern long-range dependencies and complex patterns, typical of clinical narratives.

Bi-LSTMs offer a significant advantage due to their dual-directional processing—capturing information from both past and future contexts within a sequence. This property is especially advantageous for clinical texts, which often hinge on the temporal sequence of events and interdependencies of medical terms. By integrating the Bi-LSTM layer, we achieve a more profound contextual understanding, yielding more precise and reliable classifications.

Following the Bi-LSTM, the attention layer acts as a precision tool, spotlighting the salient words pivotal for accurate classification. It addresses the potential information dilution in Bi-LSTM by applying a weighted sum to the encoded states, thus preserving valuable information. The attention weights, derived from a small dedicated neural network atop each encoded state, culminate in a single-unit output that denotes the attention weight, further refined by dense layers and a tanh activation function inspired by Bahdanau Attention.[35]

The implementation of this comprehensive encoding and attention strategy was a deliberate choice, balancing the computational overhead against the substantial gains in contextual interpretation it offers. This calculated decision underscores our dedication to innovating while remaining sensitive to the nuanced requirements of clinical text analysis.

Semantic information network. A semantic information network as shown in Figure 4(b) is used to capture domain-specific semantic information. For extracting the medical entity and their concept from the given text, a component of the scispaCy[36] NER model is utilized, and the UMLS is used as a knowledgebase for entity linker in the scispaCy component. It returns a concept unique identifier (CUI), name, definition, type unique identifier (TUI), and aliases. Embeddings are generated from the extracted UMLS semantic information for the inputted sentence, followed by Bi-LSTM and attention layers as the contextual information network. In order to provide a comprehensive semantic representation of clinical concepts, several fields are required.

Embeddings are created using the extracted UMLS semantic information. These embeddings include all the fields mentioned before: the CUI, which distinct identifies each concept; the name, which serves as a standard reference; the definition, which provides contextual understanding; the TUI, which classifies the concept within larger medical hierarchies; and aliases, which capture different synonymous expressions of the concept. Through the utilization of these multiple fields, the embeddings effectively capture both the clear identification and contextual connections of medical terminology, which are essential for precisely understanding the subtleties in clinical writing.

To enhance their representation, these embeddings are further put through a Bi-LSTM layer and an attention layer, which are analogous to the contextual information network. The incorporation of the UMLS fields guarantees that the model not only identifies particular medical items but also comprehends their wider semantic and contextual implications, which is crucial for the precise and dependable categorization of clinical narratives.

Performance evaluation metrics

To measure the merit of the algorithms, we use four statistical indicators (recall, precision, F1-score, and accuracy) for the evaluation, and the computing formulas of these

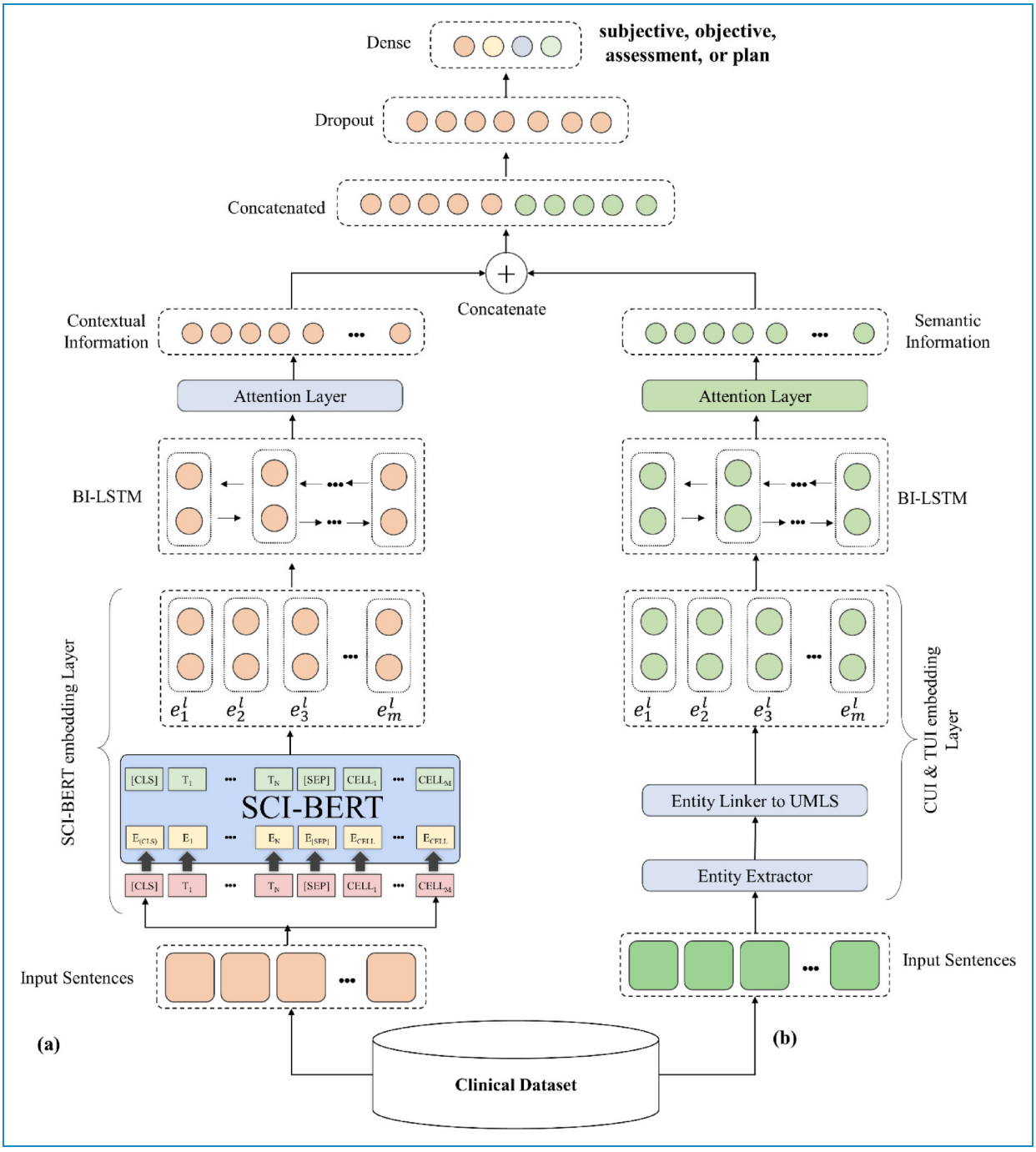

Figure 4. The proposed framework architecture shows two inputs: (a) contextual information network and (b) semantic information network, concatenated to generate multi-class output: subjective, objective, assessment, and plan.

metrics are given in equation (1).

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - Score = \frac{2(Rec * Prec)}{Rec + Prec}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

where TP: true positive, FP: false positive, TN: true negative, and FN: false negative.

Experimental results and analysis

The proposed methodology outlined earlier provides a theoretical foundation for clinical information identification and classification from unstructured clinical documents.

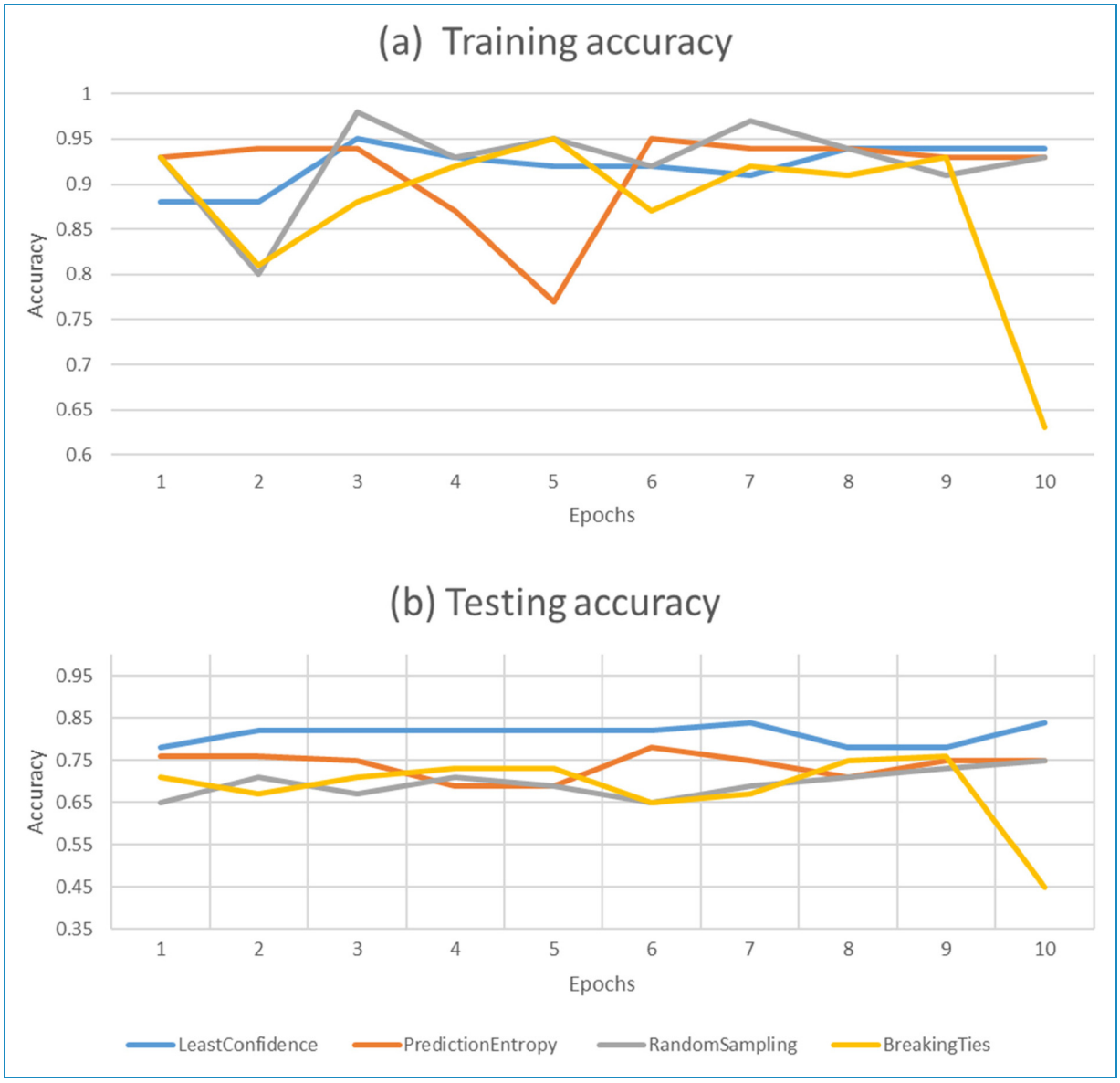

Figure 5. Performance of AL on annotations using different AL query strategies.

To construct a robust implementation of this study, it is crucial to determine the specific models and algorithms that can optimize each component individually, thereby producing high-performance intermediate results. These results can then be combined to achieve an overall optimal outcome for clinical information classification. We conducted numerous experiments to assess the effects of a rule-based approach for initial training data preparation with SOAP labels as seen in data for the AL model and evaluation of a transformer-based AL model with uncertainty-based sampling, least confidence query strategy, for annotating unlabeled clinical data with SOAP labels. Finally, we evaluated a dual attention network model incorporated SciBERT-based TL input for capturing contextual information and UMLS-based semantic enrichment (UMLS-SE) input to help capture semantic information. In this section, we have illustrated the experimental results and presented an analysis.

Active learning (AL) performance evaluation

Optimizing active learning through strategic query selection.

In the domain of AL, the efficiency of model training is often leveraged through the careful selection of data points from which the model can learn most effectively. In our recent study, we applied various AL query strategies to an annotation task on a dataset, initially comprising 243 records. These records were preannotated with a baseline

rule-based algorithm ("SOAPNotesParser"), from which we utilized the full set as the seed data to initialize our AL model. To refine the model's learning process and to maintain a manageable workload for the human annotators involved in verification, we adopted an iterative approach, selecting 290 records per iteration for model training and evaluation. For each iteration, the dataset was automatically divided into an 80/20 ratio for training and validation.

Our methodology involved a comparative analysis of four distinct query strategies within the pool-based sampling paradigm: least confidence, prediction entropy, random sampling, and breaking ties. We monitored the accuracy rates obtained by the model under each strategy across 10 iterations, aiming to ascertain the efficacy of these strategies in enhancing the model's performance.

The least confidence strategy concentrates on data points where the model has the lowest level of confidence in its predictions, usually determined by the predicted class having a probability close to 0.5. By assimilating knowledge from these ambiguous instances, the model enhances its ability to manage uncertainties. By the 10th iteration, this method attained the greatest accuracy of 94% in our testing, with minor improvements in subsequent iterations, suggesting convergence.

The prediction entropy technique chooses data points that exhibit the greatest uncertainty among all classes, employing entropy as a metric to quantify the level of unpredictability in prediction. Although rather less precise than least confidence, it outperformed random sampling by introducing diversity in the training data, therefore enabling the model to differentiate between comparable classes.

Conversely, random sampling functions as a basic reference point where data points are selected at random, without considering the uncertainty of the model. This approach yielded somewhat slower enhancements in accuracy, therefore validating the superiority of more deliberate selection techniques.

Lastly, the breaking ties strategy focuses on situations when the model encounters difficulty in selecting between two probable results. Through its emphasis on these ambiguous situations, it enhances the process of making decisions at the periphery. Although superior to random sampling, it did not achieve the same level of performance as the least confidence model.

The experimental results, depicted in Figure 5(a), illustrate the trajectory of the model's training accuracy. The least confidence strategy demonstrated superior performance, resulting in an optimal training accuracy of 94% by the 10th iteration. Notably, the accuracy plateaued between the ninth and 10th iterations, which indicated a point of convergence and served as our cue to cease further AL sample selection.

The robustness of the least confidence strategy was further validated during the testing phase, as portrayed in Figure 5(b). Here, the strategy outshone its counterparts, suggesting its greater reliability in generalizing from the AL model to unseen data. After the iterative training and testing phases, we applied the AL model to annotate the remaining records. The enriched dataset, thus augmented to encompass a total of 3146 records, will serve as a foundation for future research and applications within our AL framework.

This study affirms the value of employing judicious query strategies in AL to optimize the annotation process. The least confidence strategy has demonstrated its potential to expedite the attainment of high accuracy in model training, thereby streamlining the path toward developing more capable and efficient ML models. This approach aims to refine model predictions by focusing on cases where the model is least certain. The criteria for determining low prediction probabilities would involve threshold-based selection, where instances below a certain confidence level are reviewed. Implementing an oracle is expected to significantly improve model accuracy and reliability by ensuring only high-confidence predictions are used or by correcting mispredictions during training.

SOAP-BioMedBERT model performance valuation

In this section, we take a closer look at how well our proposed SOAP-BioMedBERT dual attention network model performs in classifying clinical text. Our goal is to see how effectively the model captures both the context and the deeper meaning of the text by combining SciBERT-based TL with UMLS-based semantic enrichment. We examined the model's performance using key metrics such as accuracy, F1 score, precision, and recall, and evaluated over several iterations of AL. We will also compare our model's performance to baseline models to highlight the improvements our approach offers. Through this analysis, we aim to show how our dual attention network can accurately sort clinical notes into the SOAP framework and demonstrate the real-world advantages of using dual attention mechanisms for clinical text classification.

Stratified k-fold and training. The dataset was subjected to a stratified k-fold ($k = 5$) cross-validation procedure to train the model and gather evaluation metrics. In this approach, the dataset is partitioned into k equally sized subsets, with each subset maintaining the same proportion of class labels as the original dataset. This method ensures that every class is appropriately represented in each fold, as demonstrated in Table 2, which outlines the distribution of classes across the folds.

During the cross-validation process, the model undergoes k iterations of training and evaluation. In each iteration, one of the k subsets is designated as the test set, and the remaining k–1 subsets serve as the training set. This strategy ensures comprehensive use of the data, with each subset getting an opportunity to be the test set exactly once. Consequently, every sample in the dataset is utilized for both training and testing purposes, promoting a thorough and balanced evaluation.

Trainable parameter optimization. To determine the most effective configuration for the model's parameters, our approach involved a systematic exploration of error rates through trial-based methods, aiming for superior accuracy in classification tasks. This involved an exhaustive search for the ideal learning rate while keeping other hyperparameters constant, to pinpoint the learning rate that minimizes loss and thus enhances the model's reliability.

In our quest to identify the most suitable optimizer for our study, we compared the performance of Adam, RMSprop (RMSP), and stochastic gradient descent (SGD) using the trial-based error approach. The outcome of this comparison favored the Adam optimizer, which demonstrated superior prediction accuracy.

Table 2. Distribution of data across five folds in stratified k-fold cross-validation.

	Subjective	Objective	Assessment	Plan	Total
Fold 1	1037	944	535	630	3146
Fold 2	1020	951	538	636	3145
Fold 3	1042	923	523	657	3145
Fold 4	1061	963	504	617	3145
Fold 5	1030	919	556	641	3146

Figure 6 illustrates the process of selecting the learning rate to optimize and assess the model. For these experiments, we utilized the sci-kit-learn library, adhering mainly to the default settings for hyperparameters. Table 3 presents a snapshot of the various hyperparameters applied to the models under study, showcasing the diversity in our experimental setup.

Each model underwent training using identical fold divisions, employing the deep learning models implementation using PyTorch framework on an NVIDIA GeForce RTX 3060 with 32 GB memory. The training process spanned 10 epochs, with hyperparameters configured to a maximum token size of 512, a batch size of 32, and a learning rate of 10e-3. Although additional epochs were explored in subsequent experiments, they did not yield significant performance improvements.

Comparative analysis between proposed (SOAP-BioMedBERT) and other BERT-based models. In our study, we initially performed experiments using traditional machine-learning models to establish a baseline for clinical text classification. Among these, the support vector machine (SVM) model was particularly noteworthy due to its versatility and effectiveness in handling high-dimensional data. Characterized by its use of a linear kernel and optimized through a meticulous process of hyperparameter tuning, the SVM model was deployed to classify clinical texts into the predefined categories of assessment, subjective, objective, and plan.

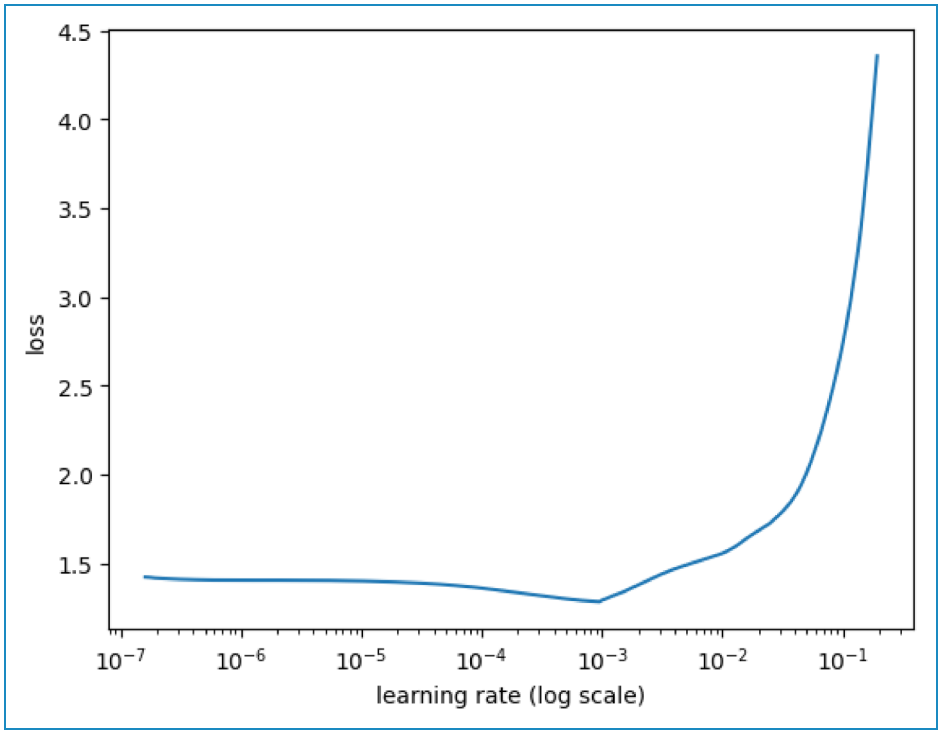

Figure 6. The learning rate selection process.

Table 3. Parameter settings of the proposed model.

Hyperparameters	Value
Max sequence length	10 k
Batch_size	32
Token size	512
Optimizer regularization	Adam
Batch normalization	True
Epochs	10
Learning rate	10e-3

The SVM model showcased promising yet mixed results in our clinical text classification study, with accuracies around 60% for all metrics, but it particularly excelled in identifying "subjective" notes. The model's lower precision and F1-scores in the "plan" category, however, pinpoint limitations that suggest the need for more sophisticated modeling techniques. The complexity of this category could be better addressed by incorporating deep learning models, which might capture the nuanced patterns of clinical data more effectively.

After the initial model training and evaluation of conventional ML algorithms, we explored advanced sophisticated deep learning algorithms. We conducted rigorous experimentation with convolutional neural networks (CNNs), recurrent neural networks (RNNs), and bidirectional long short-term memory (Bi-LSTM) networks, each employing advanced sequence word embedding techniques. These contemporary neural network architectures yielded marked improvements in terms of accuracy: CNNs reported an accuracy of 80%, RNNs achieved 93%, and Bi-LSTMs exhibited a leading accuracy of 94%.

To build upon these findings, our study then incorporated the cutting-edge BERT embeddings, a novel approach within the realm of NLP. BERT-based embeddings leverage the transformer architecture to capture complex contextual relationships within text, significantly enhancing the performance of NLP models. The adoption of BERT-based embeddings resulted in significant performance enhancements, with CNN's accuracy increasing to 85%, RNN's to 95%, and the Bi-LSTMs to an exemplary 96%. These increases represented an overall accuracy improvement ranging from 2% to 5% for the respective models as shown in Table 4. This table shows the accuracy of each model before and after the adoption of BERT-based embeddings, highlighting the improvement in percentage points. The inclusion of BERT embeddings essentially enhanced the models' ability to understand and process natural language, leading to better performance across the board.

Table 4. Improvement in neural network accuracy after integrating BERT embeddings: A comparison of CNN, RNN, and Bi-LSTM model performances before and after BERT adoption, highlighting accuracy gains.

Model	Accuracy Before BERT	Accuracy After BERT	Improvement
CNN	83%	85%	+2%
RNN	90%	95%	+5%
Bi-LSTM	91%	96%	+5%

The application of BERT-based embeddings constitutes a pivotal development given their exceptional capability to discern complex contextual interrelations within textual data. The results of this research not only affirm the formidable capabilities of transformer-based models such as BERT in processing linguistically complex datasets but also illuminate their potential to revolutionize clinical text classification methodologies. The implications of our work advocate strongly for the integration of these advanced deep learning models to effectively interpret and utilize the subtle and intricate features of clinical documentation.

Building on these preliminary findings, we embarked on a comparative study involving six preexisting BERT-based models and our newly proposed BERT-based model, SOAP-BioMedBERT. To evaluate these models rigorously, we used the results from each test iteration within the cross-validation folds to calculate the respective evaluation metrics.

Table 5 presents a comprehensive summary of the outcomes from the rigorous validation process of the six benchmarked models, including DistilBERT, BioBERT, Bio-ClinicalBERT, PubMedBERT-base, SciBERT, and our proposed SOAP-BioMedBERT model, across four categories: subjective, objective, assessment, and plan. This detailed analysis is intended to reveal the strengths and limitations of each model, including the relative improvements offered by our proposed model within the domain of clinical text classification.

The results indicate that our proposed SOAP-BioMedBERT model outperforms the other models in almost all metrics across the four categories. Specifically, it achieved the highest total weighted scores in accuracy (0.97603), precision (0.98435), recall (0.98637), and F1-score (0.98536). This showcases its superior ability to understand and categorize biomedical text with high accuracy and consistency.

Notably, SciBERT also demonstrated significant competence, especially in the subjective and objective categories, where it outpaced other models, except for our SOAP-BioMedBERT. For instance, in the subjective category, SciBERT scored a remarkable accuracy of 0.97301 and a precision of 0.97743, only slightly behind SOAP-BioMedBERT's accuracy of 0.97807 and precision of 0.98245.

Table 5. Comparative performance metrics of BERT-based models in SOAP category text classification: accuracy, precision, recall, and F1-score.

Model	Accuracy Weighted/Macro	Precision Weighted/Macro	Recall Weighted/Macro	F1-score Weighted/Macro
DistilBERT				
Subjective	0.95802	0.9638	0.97561	0.96969
Objective	0.94637	0.9750	0.97508	0.97508
Assessment	0.95491	0.9727	0.96153	0.96711
Plan	0.95530	0.9803	0.96774	0.97402
Total (raw/weighted)	0.95443	0.9725	0.97121	0.97188
BioBERT				
Subjective	0.96153	0.96385	0.97561	0.96969
Objective	0.94637	0.97706	0.97706	0.97706
Assessment	0.95491	0.97276	0.96153	0.96711
Plan	0.95530	0.98039	0.96774	0.97402
Total (raw/weighted)	0.95567	0.97330	0.97198	0.97264
Bio-ClinicalBERT				
Subjective	0.96153	0.97619	0.98795	0.98203
Objective	0.94637	0.97706	0.97706	0.97706
Assessment	0.97294	0.97276	0.96153	0.96711
Plan	0.95530	0.98039	0.96774	0.97402
Total (raw/weighted)	0.96114	0.97676	0.97544	0.97610
PubMedBERT-base				
Subjective	0.96410	0.97508	0.98798	0.98149
Objective	0.93896	0.97718	0.98010	0.97863
Assessment	0.97209	0.96911	0.95619	0.96260
Plan	0.95710	0.98066	0.96875	0.97467
Total (raw/weighted)	0.96081	0.97583	0.97583	0.97583
SciBERT				
Subjective	0.97301	0.97743	0.98918	0.98327
Objective	0.95541	0.98006	0.98794	0.98398

(continued)

Table 5. Continued.

Model	Accuracy Weighted/Macro	Precision Weighted/Macro	Recall Weighted/Macro	F1-score Weighted/Macro
Assessment	0.97472	0.97276	0.97276	0.97276
Plan	0.96142	0.98233	0.97202	0.97715
Total (raw/weighted)	0.96829	0.97846	0.98249	0.98047
Proposed (SOAP-BioMedBERT)				
Subjective	0.97807	0.98245	0.98939	0.98591
Objective	0.97280	0.98591	0.98790	0.98690
Assessment	0.97826	0.97678	0.99019	0.98344
Plan	0.97181	0.99130	0.97602	0.98360
Total (raw/weighted)	0.97603	0.98435	0.98637	0.98536

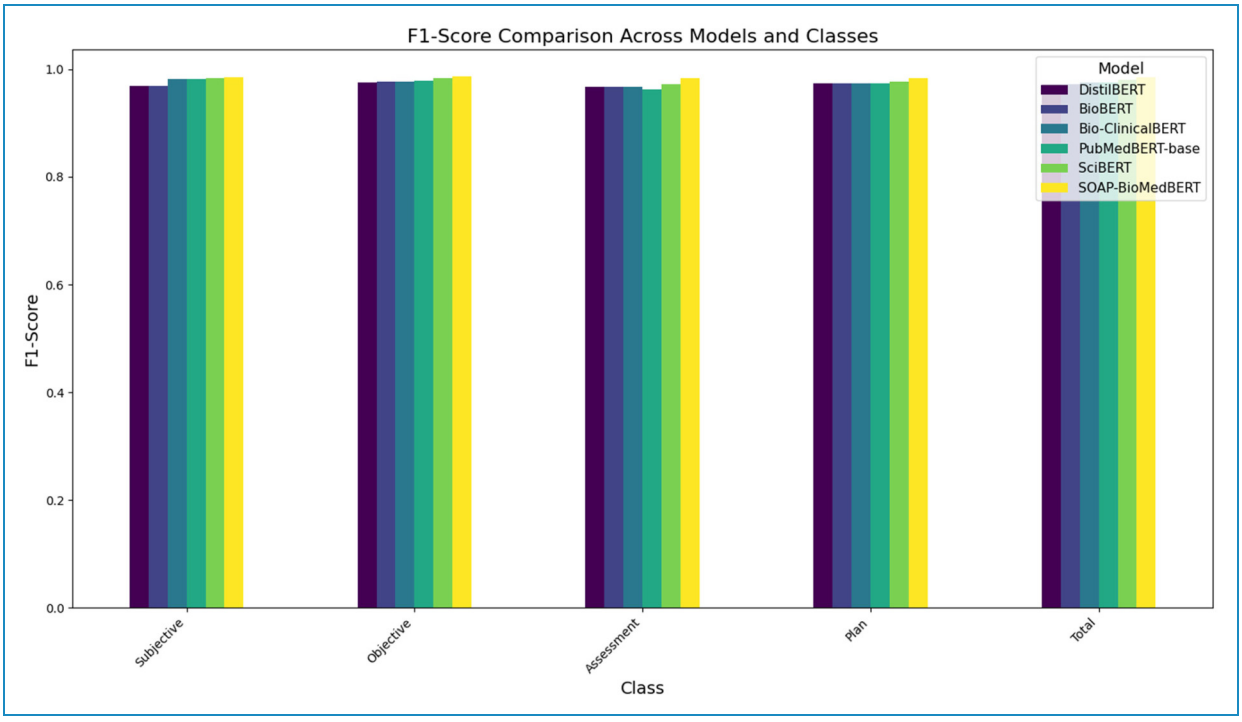

Figure 7. F1-score performance comparison of six BERT-based models across subjective, objective, assessment, plan, and total categories in biomedical text classification.

Figure 7 illustrates a comparison of F1-scores across various BERT-based models, including DistilBERT, BioBERT, Bio-ClinicalBERT, PubMedBERT-base, SciBERT, and SOAP-BioMedBERT, across five key categories: subjective, objective, assessment, plan, and total. The visualization highlights the performance disparities among these models in biomedical text classification, with SOAP-BioMedBERT emerging as the top performer across all categories. This comparison not only showcases the effectiveness of specialized models in capturing biomedical nuances but also aids in selecting the most suitable model based on a balance of computational

efficiency and accuracy for specific biomedical text analysis tasks.

The DistilBERT model, while effective, lagged behind more specialized models like BioBERT and PubMedBERT-base, reflecting the potential limitations of more generalized pretraining when applied to domain-specific tasks.

The results underscore the effectiveness of domain-specific pretraining, as evidenced by the superior performance of SOAP-BioMedBERT, which has been specifically tailored for biomedical text. This model's enhanced ability to grasp the nuances of medical literature is attributed to its training on a comprehensive corpus of biomedical texts, allowing for improved context understanding and semantic interpretation.

The slight edge of SOAP-BioMedBERT over other models can be attributed to its optimized architecture and training regimen, which was meticulously designed to capture the intricate patterns and terminologies prevalent in biomedical documents. Its performance suggests that further advancements in model architecture and training methodologies could yield even more significant improvements in text-processing capabilities for biomedical applications.

The study also highlights the importance of selecting the appropriate model for specific tasks within the biomedical domain. While general-purpose models like DistilBERT offer broad applicability, specialized models like SOAP-BioMedBERT provide the precision and accuracy necessary for high-stakes environments like healthcare and medical research.

Analysis of the best model–SOAP-BioMedBERT. The BERT model's better performance gave us the confidence to check with other embedding options. Finally, we incorporated the BERT-based embedding layer called "scibert-basevocab-uncased" together with the UMLS-based embedding layer, which produced the most excellent results of about 98% accuracies, which was better than all other configurations, and the loss was a minimum of about 1%. Our methodology incorporates a Bi-LSTM layer on top of the SciBERT model to better grasp the long-range dependencies and complex linguistic structures in clinical texts. This addition, while beneficial for model performance, introduces significant computational overhead and resource demands. The sequential nature of RNNs extends training times and requires considerable GPU resources, impacting both memory and processing power, particularly during backpropagation. To assess the trade-offs of this architectural choice, we conducted a performance-cost analysis, evaluating accuracy, precision, recall, and F1 score against training duration and GPU usage. This approach provides insights into the balance between enhanced model capabilities and the associated computational and resource implications.

The proposed model is tested on multiple points to get the desired number of epochs, and we obtained the optimal results on epoch 10 as shown in Figure 8.

In the evaluation of the model's performance over 10 epochs, two key indicators were observed: loss and accuracy, both for training and testing datasets. The upper graph showcases the training and testing loss over successive epochs. Initially, both the training and testing losses start relatively high, with the training loss demonstrating a sharp decline by the second epoch, indicating that the model is learning from the training data. The testing loss, while decreasing overall, shows fluctuations, suggesting variability in how the model generalizes to new, unseen data.

By the 10th epoch, the training loss has significantly decreased, suggesting that the model fits well with the training data. However, there is a noticeable gap between the training and testing loss, potentially indicating overfitting, as the model may not be generalizing as effectively to the testing data.

The lower graph illustrates the training and testing accuracy. Here, the training accuracy consistently improves over time, indicative of the model effectively learning and making better predictions on the training data. The testing accuracy after initial fluctuations shows an upward trend, but it does not reach the level of training accuracy by the final epoch. This again could signal overfitting, where the model's improvements are more reflective of the training data patterns rather than a generalized learning applicable to the test data.

Overall, while the model demonstrates an aptitude for learning and improving its performance on the training data, the discrepancy between training and testing metrics suggests that further tuning is required to improve generalization and prevent overfitting. Strategies such as regularization, dropout, or expanding the dataset may be considered to enhance the model's performance on unseen data.

Discussion

When comparing our work to existing studies, such as those conducted by Mowery et al.[37] and de Oliveira et al.,[5] we observe noteworthy patterns in performance as captured by the F1-scores for various classes as shown in Table 6. Our methodology yields consistently higher scores across all classes, with the "subjective" class showing a notable increase from 0.939 and 0.9477 to 0.98591. This suggests that our approach may more effectively capture the nuances of subjective information within the data.

Similarly, in the "objective" class, our F1-score of 0.98690 surpasses the previous high of 0.9566 by de Oliveira et al.,[5] indicating a stronger ability to identify and classify objective statements correctly. This improvement is critical, as objective data is often essential for drawing concrete conclusions from research findings.

The "assessment" and "plan" classes also show significant improvements in our work. The "assessment" class, which has traditionally presented challenges as indicated

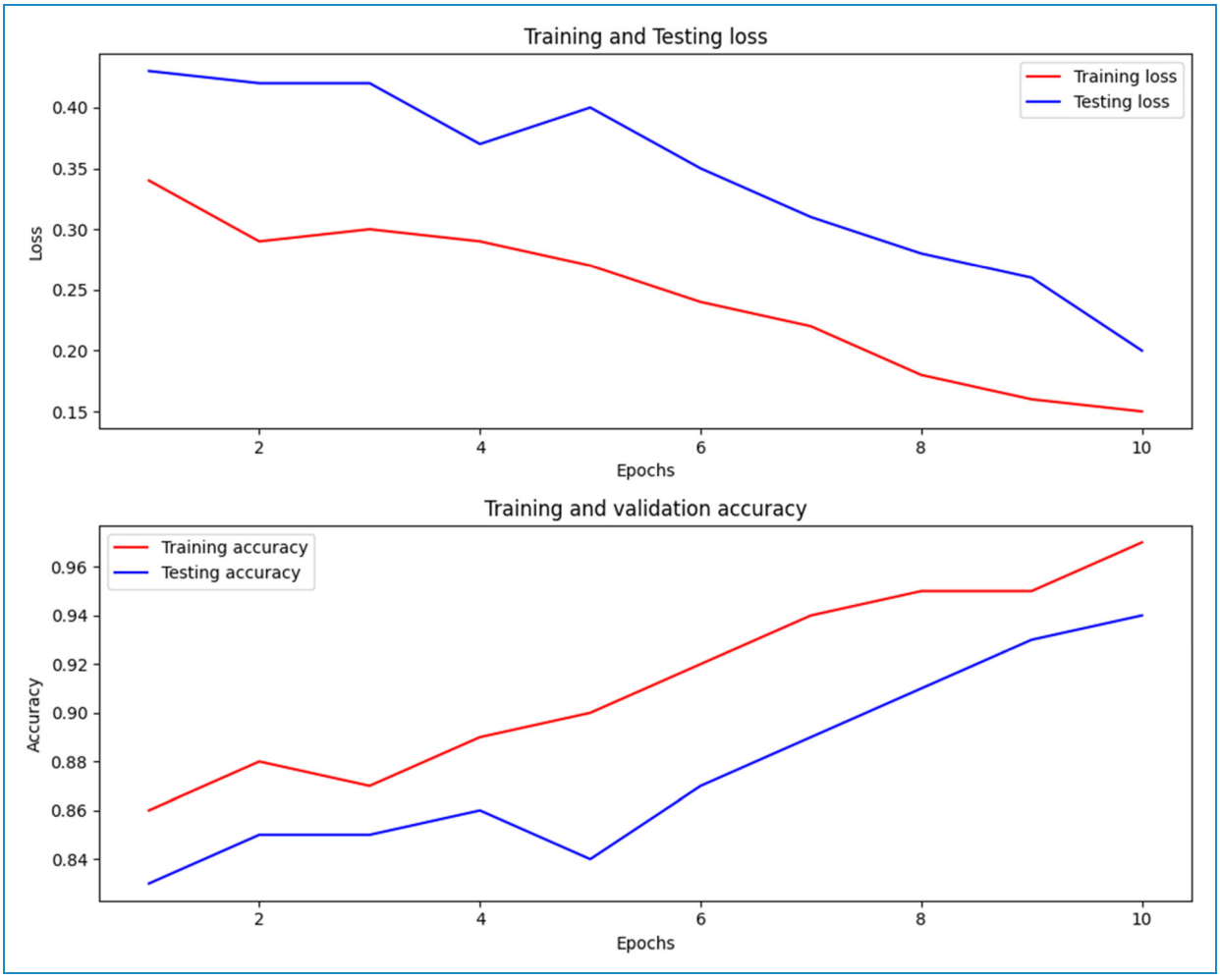

Figure 8. Proposed model accuracy on different epochs.

Table 6. Comparative analysis of F1-scores across different studies for classifying clinical notes.

Class	Mowery et al.[5] F1-Score	de Oliveira et al.[37] F1-Score	Our work (SOAP-BioMedBERT) F1-Score
Subjective	0.939	0.9477	0.98591
Objective	0.945	0.9566	0.98690
Assessment	0.757	0.7323	0.98344
Plan	0.770	0.9435	0.98360

by the relatively lower scores of 0.757 and 0.7323, sees a dramatic increase to 0.98344 in our study. This substantial enhancement suggests that our model may possess a heightened sensitivity to the key features that distinguish assessment-related content, which is crucial for medical diagnosis and treatment planning.

In the "plan" class, we see an improvement from 0.770 and 0.9435 to 0.98360, indicating our model's strength in effectively recognizing planning actions, which are imperative for the implementation of medical care.

It is important to note that while these improvements are promising, they are not solely indicative of the superiority of our model. Various factors, such as dataset composition, labeling consistency, and model architecture, can influence these results. Moreover, our model's increased performance in the "assessment" and "plan" classes, which are particularly challenging due to their predictive and prescriptive nature, may suggest a potential for our model to better understand and process complex sentence structures and semantics associated with medical decision-making.

The results underscore the Bi-LSTM layer's contribution to enhancing the model's performance on clinical text classification tasks. The incremental gains in accuracy and F1 score justify the additional computational resources, especially for

applications where precision in clinical information extraction is paramount. Nonetheless, we acknowledge the need for optimizing computational efficiency, particularly for scaling the model for larger datasets or real-time applications. Future iterations of our model will explore efficient neural network architectures and training techniques to mitigate these costs without compromising performance.

The utilization of the novel deep AL framework and the integration of transformer-based models in the research study offer significant practical implications for clinical practice, research, and patient care. Our methodology, which involves automating the annotation of clinical notes, not only reduces the manual workload required for the process but also enhances the precision and efficiency of clinical text categorization. This has direct advantages for healthcare professionals, facilitating more accurate and timely access to patient information essential for diagnosis, treatment planning, and monitoring. Additionally, our approach has the potential to advance clinical research by offering a more reliable tool for analyzing extensive datasets of patient notes, potentially revealing new insights into disease patterns, treatment results, and patient care strategies. For patients, this equates to more tailored and effective care plans, better health outcomes, and increased involvement in their care processes. By connecting advanced NLP technologies with clinical applications, our study establishes a pathway for future advancements in healthcare informatics, contributing to the overarching objective of enhancing healthcare delivery and patient outcomes.

Additionally, the proposed model can be deployed in the real-world environment to check the model's effectiveness on the real dataset. The Python code of the proposed model is provided on the GitHub link https://github.com/BioMeGiX/SOAP framework.

Case study: Applying the SOAP-based data labeling and classification framework

This case study shows how our SOAP-based data labeling and classification technique analyzes, annotates, and classifies a clinical note to demonstrate its efficacy.

Clinical note 1: Initial labeling for seed data

Clinical note with explicit headings for seed data:

Chief complaint: The patient has shortness of breath.

History of present illness: 67-year-old male with worsening shortness of breath.
Had abnormal ETT and was referred for cath. Cath revealed severe 3 vessel.
disease then referred for surgical intervention.

Physical examination: VS: 65/20 160/100 5′7″ 180 #
General: WD/WN male in NAD\\ HEENT: EOMI, PERRL, NC/AT
Neck: Supple, From, -JVD, -carotid bruits
Chest: CTAB-w/r/r
Heart : RRR -c/r/m/g
Abd: soft, NT/ND+BS
Ext: warm, well-perfused; edema; varicosities
Neuro: A\&Ox3

Discharge diagnosis: Rectal bleeding from inferior mesenteric artery tributaries supplying sigmoid colon.

Preprocessing and labeling using SOAPNotesParser. Using the "SOAPNotesParser" algorithm, each sentence is labeled according to the SOAP framework:

- *Subjective (chief complaint).* Shortness of breath.
  Explanation: This section reflects the patient's primary concern, fitting the subjective category.
- *Objective (physical examination).* "VS: 65/20, 160/100, 5'7," 180 lbs. General: WD/WN male in NAD\ HEENT: EOMI, PERRL, NC/AT\ Chest: CTAB -w/r/r Heart: RRR -c/r/m/g Abd: soft, NT/ND+BS\ Ext: warm, well-perfused, -edema, -varicosities."
  Explanation: This sentence captures objective clinical observations, fitting the objective category.
- *Assessment (history of present illness).* 67-year-old male with worsening shortness of breath. Had abnormal ETT and was referred for cath. Cath revealed severe 3 vessel disease, then referred for surgical intervention.
  Explanation: This section includes the physician's diagnostic assessment, fitting into the assessment category.
- *Plan (discharge diagnosis).* Rectal bleeding from inferior mesenteric artery tributaries supplying the sigmoid colon.
  Explanation: The discharge diagnosis and treatment outline fall under the plan category.

These labeled sentences form the initial seed dataset that will train the AL model.

Clinical note 2: Active learning process

Clinical note: The patient complains of persistent headaches and blurred vision over the past few days. Physical examination shows no neurological deficits, but blood pressure is significantly elevated. The assessment is that the patient may be experiencing hypertensive crisis. Plan: recommend lifestyle changes, initiate antihypertensive therapy, and schedule follow-up.

Preprocessing. The clinical note is segmented into the following sentences:

- "Patient complains of persistent headaches and blurred vision over the past few days."
- "Physical examination shows no neurological deficits, but blood pressure is significantly elevated."
- "The assessment is that the patient may be experiencing hypertensive crisis."
- "Plan: recommend lifestyle changes, initiate antihypertensive therapy, and schedule follow-up."

Active learning strategy. The AL model, trained on the initial seed data, uses the least confidence strategy to select sentences with low prediction confidence for manual annotation. In this note, the model shows low confidence in classifying the sentence:

"Patient complains of persistent headaches and blurred vision over the past few days."

This sentence is first labeled by the proposed AL model, then reviewed by a human expert to ensure accurate classification. It is then added back into the training dataset. This process continues iteratively, refining the model's ability to classify clinical notes accurately.

Annotation by model and verified by oracle.

- *Subjective*: "Patient complains of persistent headaches and blurred vision over the past few days."
  Explanation: This sentence is identified as patient-reported symptoms, fitting the subjective category.
- *Objective*: "Physical examination shows no neurological deficits, but blood pressure is significantly elevated."
  Explanation: This sentence describes observations made during the physical examination, fitting into the objective category.
- *Assessment*: "The assessment is that the patient may be experiencing a hypertensive crisis."
  Explanation: This sentence provides the clinician's diagnostic impression, aligning with the assessment category.
- *Plan*: "Plan: recommend lifestyle changes, initiate antihypertensive therapy, and schedule follow-up."
  Explanation: This sentence outlines treatment and follow-up actions, categorizing it under the plan section.

These additional labeled instances enhance the training dataset, improving the model's accuracy and reliability.

Clinical note 3: Final testing of the model

Clinical note: The patient describes feeling extremely fatigued and having had a persistent dry cough for the last two weeks. Physical examination indicates decreased breath sounds in the right lung. The assessment is that the patient may have pneumonia. Plan: Start antibiotics and order a chest X-ray.

Preprocessing. The clinical note is segmented into the following sentences:

- "Patient describes feeling extremely fatigued and having a persistent dry cough for the last two weeks."
- "Physical examination indicates decreased breath sounds in the right lung."
- "The assessment is that the patient may have pneumonia."
- "Plan: Start antibiotics and order a chest X-ray."

Final model classification. Using the trained deep learning model, which has been fine-tuned through the AL process, each sentence is classified into the appropriate SOAP category:

- *Subjective*: "Patient describes feeling extremely fatigued and having a persistent dry cough for the last two weeks."
- *Objective*: "Physical examination indicates decreased breath sounds in the right lung."
- *Assessment*: "The assessment is that the patient may have pneumonia."
- *Plan*: "Plan: Start antibiotics and order a chest X-ray."

Conclusions, limitations, and future works

The vast availability of unstructured clinical data offers an opportunity to extract meaningful information for the applications that support the process of clinical decision-making. However, extracting the relevant information from unstructured text into a clinically useful format is a big challenge. Therefore, this work targeted this aspect of information extraction into a well-known protocol (SOAP) used as an information container. The clinical text in the form of SOAP structure enhances information readability, and the individual sentences, that is, subjective, objective, assessment, and plan, can be used in other add-on applications such as clinical decision support systems. Additionally, it helps organizations develop multiple individualistic systems such as diagnostic, treatment, and prognostic by utilizing the relevant SOAP section.

Despite the promising results, our study has a few limitations. Firstly, the performance of the proposed model heavily relies on the availability of high-quality labeled data. The quality and accuracy of the annotations can significantly impact the classification performance. Additionally, the generalizability of our model may be limited to the specific context of the i2b2 dataset and may require adaptation when applied to other datasets or clinical settings. Furthermore, the proposed methodology assumes the SOAP framework, and its effectiveness may vary

when applied to different medical protocols or classification tasks.

In future research, an interesting direction to explore is the incorporation of prompt engineering techniques based on large language models (LLMs). LLM-based prompt engineering has shown promising results in improving the performance of language models on various NLP tasks. Integrating LLM techniques into the proposed methodology for clinical text annotation and classification could yield further improvements.

One potential approach is to leverage LLMs to generate informative prompts for AL. These generated prompts can help direct the annotation effort toward more informative samples, enhancing the efficiency and effectiveness of the AL model. Furthermore, LLMs can be used to refine and adapt the transformer-based model for the specific domain of clinical notes and SOAP classification. Prompt-based fine-tuning techniques could be explored to optimize the models' performance on the SOAP classification task.

Additionally, exploring the combination of AL and LLM-based prompt engineering can lead to enhanced annotation quality and model performance. By leveraging the contextual knowledge and capabilities of LLMs, models can better understand the clinical context and improve the accuracy of the generated annotations during the AL iterations.

Overall, incorporating LLM-based prompt engineering techniques into the proposed methodology has the potential to further advance the field of clinical text annotation and classification. It can enhance the efficiency, accuracy, and generalizability of models, making them more robust in handling variations in clinical notes and improving their performance on the SOAP framework.

Contributorship: MA, JH, and AB contributed to conceptualization. MA, JH, and AB contributed to data curation. MA, JH, and AB contributed to investigation. JH, MA, and AB contributed to methodology. JH, MA, and AB contributed to writing original draft. All authors have read, reviewed, and approved the final manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This study did not require ethics committee review and approval.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2024-RS-2020-II201489) and was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program(IITP-2024-RS-2023-00259004) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2022-II220078, Explainable Logical Reasoning for Medical Knowledge Generation), (RS-2017-II170655, Lean UX core technology and platform for any digital artifacts UX evaluation).

Guarantor: Muhammad Afzal.

Patient consent statement: This research utilized a publicly available dataset that did not contain any directly identifiable patient information. Therefore, informed patient consent was not required for this study.

ORCID iDs: Jamil Hussain https://orcid.org/0000-0003-3862-8787
Asim Abbas https://orcid.org/0000-0001-6374-0397

References

1. Yao L, Mao C and Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Med Inform Decis Mak* 2019; 19: 31–39.
2. Liang J, Tsou C-H and Poddar A. A novel system for extractive clinical note summarization using EHR data. In: *Proceedings of the 2nd clinical natural language processing workshop*. Minneapolis, MN: Association for Computational Linguistics, 2019, pp.46–54.
3. Li I, Yasunaga M, Nuzumlalı MY, et al. A neural topic-attention model for medical term abbreviation disambiguation. *ArXiv-> Computer Science > Computation and Language* 2019. Epub ahead of print 2019. DOI: 10.48550/ARXIV.1910.14076
4. Seyedmostafa S, Miotto R, Dudley JT, et al. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019; 7: e12239.
5. Mowery D, Wiebe J, Visweswaran S, et al. Building an automated SOAP classifier for emergency department reports. *J Biomed Inform* 2012; 45: 71–81.
6. Weng W-H, Wagholikar KB, McCray AT, et al. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak* 2017; 17: 1–13.
7. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018; 77: 34–49.
8. Wang Y, Sohn S, Liu S, et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 2019; 19: 1–13.
9. Irena S and Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform* 2020; 8: e17984.
10. Kholghi M, Sitbon L, Zuccon G, et al. Active learning reduces annotation time for clinical concept extraction. *Int J Med Inform* 2017; 106: 25–31.
11. Searle T, Kraljevic Z, Bendayan R, et al. MedCATTrainer: a biomedical free text annotation interface with active learning and research use case specific customisation. 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): System Demonstrations, pp.139–144. Hong Kong, China: Association for Computational Linguistics, 2019. Epub ahead of print 2019. DOI: 10.48550/ARXIV.1907.07322
12. Schuyler PL, Hole WT, Tuttle MS, et al. The UMLS metathesaurus: representing different views of biomedical concepts. *Bull Med Libr Assoc* 1993; 81: 217.
13. Whetzel PL, Noy NF, Shah NH, et al. Bioportal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res* 2011; 39: W541–W545.
14. Khan J and Lee Y-K. LeSSA: a unified framework based on lexicons and semi-supervised learning approaches for textual sentiment classification. *Applied Sciences* 2019; 9: 5562. Epub ahead of print 2019. DOI: 10.3390/app9245562
15. Beltagy I, Lo K and Cohan A. SciBERT: a pretrained language model for scientific text. *ArXiv–> Computer Science > Computation and Language* 2019. Epub ahead of print 2019. DOI: 10.48550/ARXIV.1903.10676
16. Biomedical Informatics (DBMI) at Harvard Medical D. i2b2: informatics for integrating biology and the bedside. https://www.i2b2.org/NLP/DataSets/Main.php
17. Hussain M, Satti FA, Hussain J, et al. A practical approach towards causality mining in clinical text using active transfer learning. *J Biomed Inform* 2021; 123: 103932.
18. Church KW. Word2Vec. *Nat Lang Eng* 2017; 23: 155–162.
19. An N, Xiao Y, Yuan J, et al. Extracting causal relations from the literature with word vector mapping. *Comput Biol Med* 2019; 115: 103524.
20. Li M, Scaiano M, El Emam K, et al. Efficient active learning for electronic medical record de-identification. *AMIA Summits on Translational Science Proceedings* 2019; 2019: 462.
21. Tomanek K and Hahn U. Annotation time stamps—temporal metadata from the linguistic annotation process. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). May 17-23, 2010, Valletta, Malta: European Language Resources Association (ELRA), 2010. http://www.lrec-conf.org/proceedings/lrec2010/pdf/652_Paper.pdf
22. Yukun Chen B, Denny JC, Hua Xu M, et al. Active learning for named entity recognition in clinical text. *J Biomed Inform* 2015; 58: 11–18. DOI: 10.1016/j.jbi.2015.09.010.
23. Zhou S, Chen Q and Wang X. Active deep learning method for semi-supervised sentiment classification. *Neurocomputing* 2013; 120: 536–546.
24. Hajmohammadi MS, Ibrahim R, Selamat A, et al. Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Inf Sci (N Y)* 2015; 317: 67–77.
25. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. 2013. Epub ahead of print 2013. DOI: 10.48550/ARXIV.1301.3781
26. Pennington J, Socher R and Manning CD. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). October 25-29, 2014, pp. 1532–1543. Doha, Qatar.
27. Devlin J, Chang M-W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv–> Computer Science > Computation and Language* 2018. Epub ahead of print 2018. DOI: 10.48550/ARXIV.1810.04805
28. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2019. Epub ahead of print September 2019. DOI: 10.1093/bioinformatics/btz682
29. Huang K, Altosaar J and Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. *ArXiv–> Computer Science > Computation and Language* 2019. Epub ahead of print 2019. DOI: 10.48550/ARXIV.1904.05342
30. Eriksen MB and Frandsen TF. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *J Med Libr Assoc* 2018; 106: 420.
31. Abbas A, Hussain J, Afzal M, et al. Explicit and implicit section identification from clinical discharge summaries. In: 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM). 3–5 Jan 2022, pp. 1–8. Seoul, South Korea.
32. Abbas A, Afzal M, Hussain J, et al. Clinical Concept extraction with lexical semantics to support automatic annotation. *Int J Environ Res Public Health* 2021; 18: 10564. Epub ahead of print 2021. DOI: 10.3390/ijerph182010564
33. Schröder C, Müller L, Niekler A, et al. Small-text: active learning for text classification in Python. *ArXiv–> Computer Science > Machine Learning* 2021. Epub ahead of print 2021. DOI: 10.48550/ARXIV.2107.10314
34. Lewis DD and Gale WA. A sequential algorithm for training text classifiers. In: SIGIR'94. July 3–6, 1994, pp. 3–12. Dublin Ireland.
35. Bahdanau D, Cho K and Bengio Y. Neural machine translation by jointly learning to align and translate. *ArXiv-> Computer Science > Computation and Language* 2014. Epub ahead of print 2014. DOI: 10.48550/ARXIV.1409.0473
36. Neumann M, King D, Beltagy I, et al. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp.319-327, Florence, Italy. Association for Computational Linguistics, 2019.
37. de Oliveira JM, Antunes RS and da Costa CA. SOAP classifier for free-text clinical notes with domain-specific pre-trained language models. *Expert Syst Appl* 2024; 245: 123046.