BMBAD BJESR Mühendislik Bilimleri ve Araştırmaları Dergisi
Journal of Engineering Sciences and Researches

Medical Text Classification Using Semisupervised Learning and Bert-Based Models

Yarı Denetimli Öğrenme ve Bert Tabanlı Modeller Kullanılarak Tıbbi Metin Sınıflandırma

[1]Fatih SOYGAZİ, [2]Damla OĞUZ

[1]Aydın Adnan Menderes University, Faculty of Engineering, Department of Computer Engineering, Aydın, Türkiye
[2]İzmir Institute of Technology, Faculty of Engineering, Department of Computer Engineering, İzmir, Türkiye

fatihsoygazi@adu.edu.tr, damlaoguz@iyte.edu.tr

Araştırma Makalesi/Research Article

ARTICLE INFO

Article history

Received : 6 December 2024
Accepted : 13 February 2025

Keywords:

BioBERT, ClinicalBERT, Clinical Text Classification, Data Augmentation, Voting Mechanisms

ABSTRACT

Medical text classification organizes complex medical texts, facing challenges like insufficient training data. This paper proposes a novel method for categorizing medical texts based on a dataset of health problem abstracts and their labels. We applied data representation techniques to our labeled dataset and employed various machine learning algorithms for text classification. Initial results were unsatisfactory due to limited labeled data. To enhance this, we applied data augmentation techniques using an unlabeled dataset, utilizing BERT-based models (BioBERT, ClinicalBERT) to enrich the labeled data. Different voting mechanisms, namely hard voting and soft voting were employed to validate and add new labeled records to the dataset. After augmenting the labeled data, machine learning algorithms were re-applied. The results demonstrated that our approach significantly improves the performance of medical text classification, effectively addressing the challenges posed by limited labeled data and enhancing overall accuracy.

MAKALE BİLGİSİ

Makale Tarihleri

Gönderim : 6 Aralık 2024
Kabul : 13 Şubat 2025

Anahtar Kelimeler:

BioBERT, ClinicalBERT, Klinik Metin Sınıflandırması, Veri Artırma, Oylama Mekanizmaları

ÖZET

Tıbbi metin sınıflandırması, yetersiz eğitim verisi gibi zorluklarla karşılaşarak karmaşık tıbbi metinleri düzenlemektedir. Bu çalışma, sağlık sorunları özetleri ve etiketleri içeren bir veri setine dayanarak tıbbi metinleri sınıflandırmak için yeni bir yöntem önermektedir. Etiketli veri setimize veri temsil teknikleri uyguladık ve metin sınıflandırması için çeşitli makine öğrenmesi algoritmaları kullandık. İlk sonuçlar, sınırlı etiketli veriler nedeniyle yeterli bulunmamıştır. Bunu geliştirmek için, etiketli verileri zenginleştirmek amacıyla etiketlenmemiş bir veri seti kullanarak veri artırma teknikleri uyguladık; bu süreçte BERT tabanlı modeller (BioBERT, ClinicalBERT) kullanılmıştır. Yeni etiketli kayıtları doğrulamak ve veri setine eklemek için çoğunluk oylama ve ağırlıklı çoğunluk oylama gibi farklı oylama mekanizmaları kullanılmıştır. Etiketli verileri artırdıktan sonra, makine öğrenmesi algoritmalarını yeniden uygulanmıştır. Sonuçlar, yaklaşımımızın tıbbi metin sınıflandırmasının performansını önemli ölçüde artırdığını, sınırlı etiketli verilerin getirdiği zorlukları etkili bir şekilde ele aldığını ve genel doğruluğu artırdığını göstermiştir.

ORCID ID: [1]0000-0001-8426-2283
[2]0000-0001-6556-7444

1. INTRODUCTION

Medical text classification is a specialized area of text classification that deals with organizing and categorizing medical texts. These texts contain complex medical language and measurements, which can make classification more challenging due to high-dimensionality and sparsity of data. With the growing volume of digital documents, automatic text classification has emerged as a significant research area. This importance extends to the medical field, particularly with electronic health record (EHR) data. As the amount of medical data grows, the goal of medical text classification becomes crucial: to accurately classify medical records, reports, and other relevant texts into specific categories or classes based on their content.

Collecting medical documents, however, presents challenges due to ethical and privacy concerns. To address this, augmenting missing medical data based on existing information is crucial. Data augmentation directly contributes to the success of machine learning and deep learning applications in the medical field, helping to bridge data gaps and support better outcomes in medical research. This study specifically aims to address the critical challenge of limited labeled data in medical text classification, which restricts the effective training of supervised machine learning models. By leveraging semi-supervised learning techniques and clinical text-oriented BERT models, the proposed method significantly increases the volume of labeled data, thereby effectively mitigating the adverse effects of data scarcity in medical text classification. The approach enables the integration of domain-specific embeddings with traditional ML algorithms, improving the performance and robustness of classification tasks.

With the recent advancements in Natural Language Processing (NLP) and the success of transformer-based deep learning (DL) models such as BERT, there has been a growing interest in applying BERT-based approaches for medical text classification. But one of the major problems in these kinds of supervised learning models is the lack of labeled data. Labeled data is crucial for training supervised models to make accurate predictions. However, obtaining labeled data can be a challenging and time-consuming process, particularly in domains like medicine where human annotation is required. In the medical domain, accurately labeling unlabeled data demands a high level of expertise. Therefore, employing an automatic labeling approach in this context would prove immensely beneficial.

The problem of lack of labeled data can have several negative impacts on DL. First, models may underfit due to the limited amount of available labeled data, resulting in poor accuracy on both training and test sets. Second, models may not be able to capture important patterns and relationships in the data, leading to suboptimal performance on specific tasks. Finally, without enough labeled data, it can be difficult to fine-tune pre-trained models effectively, which can limit their performance on downstream tasks. Therefore, this work focuses on increasing the small amount of labeled data using the variants of various types of transformer models.

In this study, medical text classification results were initially obtained using traditional machine learning (ML) methods. A medical text classification study, considering medical specialties as target values, was conducted using 14,438 labeled and 14,442 unlabeled medical texts obtained from Kaggle [1]. Results such as accuracy, precision, recall, and F1 score were obtained through studies conducted using various ML algorithms including K-nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM).

For improved results, we considered augmenting the unlabeled text data essential [2]. We increased the labeled text volume by separately retraining the models with BioBERT [3] and ClinicalBERT [4]. We then integrated the newly labeled texts into the initial training dataset. Additionally, we improved the evaluation results for the unlabeled dataset using soft-voting and hard-voting. The best results were achieved using soft voting in combination with Random Forest (RF), yielding precision, recall, and F1 score values of 0.90.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Section 3 outlines the methodology. Section 4 presents the results. Section 5 discusses similar studies covering augmentation and voting mechanisms. Lastly, Section 6 concludes the paper.

2. RELATED WORK

The paper focuses on two main topics: text classification and data augmentation. For text classification, traditional machine learning methods, ensemble techniques, and deep learning approaches can be utilized. A sufficient amount of labeled data is essential for successful text classification. If the results fall below the expected performance values, it is often due to the available labeled data not adequately representing the model. In such cases, text classification results can potentially be improved through data augmentation methods. The most effective method depends on the specific task and the dataset being used.

Text classification is a specific application of ML that categorizes text documents into predefined classes based on their content. It has been used in many different applications, including the medical domain where medical text data is classified. This is known as medical text classification and considered as a special case of text classification [5, 6]. The state-of-the-art approaches for medical text classification can be classified into traditional ML-based methods, DL-based methods, transfer learning, hybrid methods, and ensemble methods [7, 8]. Traditional machine learning (ML)-based methods are effective for small datasets and are relatively easy to interpret. However, they have limited scalability and require manual feature extraction, which can be a significant drawback for complex tasks. Deep learning (DL)-based methods, on the other hand, are capable of handling complex data and can automatically learn features, making them highly effective for large-scale problems. Nonetheless, they require

substantial amounts of labeled data and are computationally intensive, which can pose challenges. Transfer learning leverages pre-trained models to reduce training time, offering a valuable advantage, particularly in domains with limited labeled data. However, its effectiveness is constrained by the quality of the pre-trained models, and a transfer gap may exist when applied to dissimilar tasks. Ensemble methods combine predictions from multiple models to improve accuracy and reduce variance, making them highly robust. However, they come with high computational costs and the risk of overfitting if the models are not carefully chosen. Lastly, hybrid methods aim to combine the strengths of ML and DL techniques, often achieving improved performance by using DL for feature extraction and ML for classification. Despite their potential, they require significant amounts of labeled data and add complexity to the workflow.

ML methods use algorithms, such as KNN, DT, RF, and SVM, to learn patterns between words and predefined categories from labeled medical text data. Textual data must be transformed into numerical features using methods such as bag-of-words or word embeddings. Then, the mentioned ML algorithms use these numerical representations as inputs. Although ML-based methods can be effective with a small amount of labeled data, they may not be sufficient on their own for large datasets [9].

DL-based methods can classify medical text data using neural networks such as recurrent neural networks (RNNs). They have the ability to automatically learn features from the data. Unlike ML-based methods, DL can be sufficient on its own for large datasets [10, 11]. Moreover, DL can handle more complex datasets and also can be customized to consider specific characteristics of the medical domain [11, 12]. This can be achieved through the use of appropriate training datasets or other techniques, including transfer learning [13], which involves utilizing a pre-trained model to solve a new problem. Transfer learning can also be used with ML-based models to have pre-trained models and adapt them to new tasks [14]. Since RNNs cannot handle long-range dependencies and cannot benefit from parallelization, new approaches are proposed such as transformer models [15]. One of the most widely used and well-known transformer models is BERT (Bidirectional Encoder Representations from Transformers) [16], which is trained on large amounts of text data. It can be fine-tuned on smaller datasets for specific tasks including text classification. BioBERT ((Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [3] and ClinicalBERT [4] are variants of BERT that have been specifically trained on clinical and biomedical text data, respectively to improve the performance on the tasks in medical domain. BioBERT has been used for various tasks such as biomedical named entity recognition [17] and relation extraction [18, 19], while ClinicalBERT has been used for tasks including clinical named entity recognition [20].

Ensemble methods are another used approach for medical text classification which combines the predictions of multiple models [21]. In other words, a set of models on the same dataset are needed for training and the predictions of them are combined in different ways such as using majority voting and weighted voting.

Hybrid methods that combine ML and DL techniques are also being used in medical text data classification. In typical hybrid methods, a DL model, such as RNN is used to extract features from the data, and then a traditional ML algorithm such as SVM is performed for the medical text classification. In other words, the output of the DL model is used as an input for the ML algorithm. Although hybrid methods can have better performance than using either approach alone, they need larger amounts of labeled training data as DL-based methods.

Combining several techniques can lead to better performance [9]. Therefore, data augmentation should be employed when the dataset is limited [2, 22]. This becomes particularly relevant in the biomedical field, where obtaining large, balanced datasets can be challenging due to privacy concerns and data availability. Therefore, researchers have turned to data augmentation as an effective method to address these limitations.

There are various papers that explore data augmentation with BioBERT and ClinicalBERT. Lu et al. [23] explore textual data augmentation with ClinicalBERT for predicting patient outcomes, focusing specifically on patient readmission. Their experiments demonstrate that ClinicalBERT benefits significantly from this strategy, outperforming other augmentation methods. In this study, ClinicalBERT is integrated into the MedAug framework, which uses a fine-tuned GPT-2 model to generate additional training data, addressing issues of data scarcity and class imbalance. To manage noise, a teacher-student framework guides the student model, trained on both original and synthetic data, by enforcing knowledge consistency. While this approach improves model performance, challenges such as high computational costs, scalability limitations, and residual noise remain. Nonetheless, MedAug proves to be an effective tool for enhancing text-based predictive tasks in healthcare. Erdengasileng et al. [24] explore the use of BioBERT and ClinicalBERT, combined with data augmentation and ensemble learning strategies, for biomedical information extraction and document classification tasks. The study introduces key data augmentation techniques, such as modifications to drug/chemical–protein interactions, chemical entity recognition, and medication extraction from tweets. These strategies aim to enhance the diversity and robustness of training data, ultimately improving model performance for specific biomedical applications. By leveraging pre-trained models, data augmentation, and ensemble learning, the study achieves significant advancements in performance across multiple tracks of the BioCreative Challenge VII, highlighting the practical effectiveness of these methods. However, the approach has certain limitations. It heavily relies on computationally intensive ensemble models and pre-trained architectures, making it resource demanding. Additionally, the applicability of the methods may be restricted in real-world scenarios where domain-specific pre-trained models or sufficient labeled data are unavailable. Zhang et al. [25] explore data augmentation in medical specialty classification using BioBERT as the classifier model. Their primary aim is to tackle challenges such as insufficient and imbalanced medical text data, which often hinder the performance of classification models. By fine-tuning

BioBERT with augmented data, the study achieves significant performance improvements compared to other models like CNN, LSTM, and standard BERT. The study introduces a novel Semi-Adversarial Data Augmentation (SemiADA) technique, paired with probabilistic information recalculation, to address these challenges. This approach boosts model performance by 15.1% in accuracy and 14.7% in F1 score, making it particularly effective for datasets with multiple classes. SemiADA is not only cost-effective but also enhances the model's robustness and generalization capabilities. However, the method has certain limitations. It is computationally intensive, especially during the data augmentation and training phases. Additionally, while the probabilistic information layer improves classification for underrepresented categories, the approach's scalability and efficiency may decrease as data complexity or size increases, requiring further optimization for broader real-world applications.

3. METHODOLOGY

3.1. Dataset

The “Medical Text” dataset [1] comprises medical abstracts that describe current conditions of patients. The patient-oriented text is categorized into five different classes: digestive system diseases, cardiovascular diseases, neoplasms, nervous system diseases, and general pathological conditions. The dataset includes a total of 14438 labeled patient records and 14442 unlabeled ones. The primary objective of this dataset is to utilize the labeled records for training while predicting the classification of the unlabeled records. This dataset contributes to assisting doctors by enabling them to identify salient information regarding each patient’s condition, thereby facilitating a more informed diagnosis. To illustrate the dataset's structure, Table 1 presents the distribution of the labeled patient records across various disease categories.

Table 1. Class distribution of medical text dataset.

Index	Label	Support (Count)
1	digestive system diseases	3310
2	cardiovascular diseases	1476
3	neoplasms	1876
4	nervous system diseases	3023
5	general pathological conditions	4753

3.2. Data Preprocessing and Representation

Terminological words in the medical domain are mostly long. Hence, words with less than five letters are initially removed from the dataset to eliminate meaningless or incorrectly typed words. The contextual information in the medical domain can generally be captured by specific words in a medical discipline. Our aim is to focus on these words while ignoring punctuation marks. After cleaning the text, we generate a Bag-of-Words (BoW) model to analyze it. This model represents a piece of text as a collection of its constituent words, regardless of the order of the words. Extracting relevant features from medical text data can aid in classification, and using the BoW model for a simple representation before classification is a good starting point. However, the large number of words, including critically important medical terms, must be represented in a more processable manner. Therefore, a CSR (Compressed Sparse Row) matrix is created to build a sparse matrix from the list of abstracts in the dataset. CSR matrices allow efficient memory access and arithmetic operations on sparse matrices, especially when the matrices are very large and sparse. When collecting data in a broad range of medical texts, using CSR matrices is an effective method of data storage. The CSR matrix stores the frequency of words in each textual document.
IDF (Inverse Document Frequency) is a numerical measure that reflects how important a term is to a collection of documents. It is commonly used in text mining, NLP, and information retrieval. The idea behind IDF is that some terms are more important than others in a document collection, and these terms should be given more weight in information retrieval systems. IDF calculates the weight of a term by dividing the total number of documents in the collection by the number of documents containing the term. The resulting value is then logarithmically scaled to reduce the effect of high document frequency terms. When the IDF value of a term is larger than the other one means that this is relatively an important term in the collection, as it appears in a relatively small proportion of documents. To understand the importance of the terms in the medical text collection, IDF values of each term are calculated in the term list of CSR matrix.

3.3. Data Augmentation via BERT-based Learning

Studies conducted with a small amount of labeled data may not yield successful results in machine learning models and the fact that the available data may not provide sufficient information about the relevant label. In this regard, if there is unlabeled data available, labeling it using the relevant labels is crucial to improving success. To address this challenge, we initially obtained baseline results by evaluating the performance on the available labeled data. Since the dataset obtained from Kaggle also contains unlabeled data, the focus was first placed on labeling this data using various machine learning methods. We used semi-supervised learning to apply this approach. Semi-supervised learning is a machine learning approach that falls between supervised and unsupervised learning. It uses both labeled and unlabeled data to improve learning performance, particularly when labeled data is scarce or

expensive to obtain. Hence, it was anticipated that the results could be improved by examining semi-supervised learning utilizing BERT models specific to the medical domain, BioBERT and ClinicalBERT.
Since our work involves a small amount of labeled data, the main challenge is to enhance the labeled dataset we have. Given that the context pertains to patient health records, it is essential to leverage relevant medical information corpora. BioBERT and ClinicalBERT can provide contextual insights into these medical records, allowing for the classification of patients' biological or clinical backgrounds into various disease categories. However, the classification outputs from each BERT model may yield different results. Therefore, each model should be integrated in a combined manner. To achieve this, soft and hard voting approaches have been applied to aggregate the information provided by each model. The output of the augmented data for the best-performing model is presented in Table 2.
As the dataset involves a small amount of labeled data for each category, we decided to apply BioBERT and ClinicalBERT to increase the number of records in each category. BioBERT [3] is a variant of the BERT model specifically pre-trained for tasks in the biomedical domain. It involves the datasets taken from PubMed (a massive database containing millions of biomedical abstracts), over 4.5 billion words of abstracts, and PMC (PubMed Central), a free full-text archive of biomedical and life sciences including over 13.5 billion words of full-text articles. BioBERT significantly improves upon traditional NLP models in medical and biological contexts, and it assists to understand domain-specific terminology and context on clinical documentation, medical research, and healthcare data analysis. On the other hand, ClinicalBERT [4] is a specifically adapted BERT model for understanding and processing clinical text data such as electronic health records (EHRs). It enhances the standard BERT model's ability to interpret medical jargon and clinical context, which is critical in healthcare applications like patient record analysis, clinical decision support, and medical document classification. ClinicalBERT involves text from MIMIC-III (Medical Information Mart for Intensive Care) dataset (a large, publicly available dataset containing de-identified health records of over 40,000 patients admitted to critical care units including discharge summaries, nursing notes, radiology reports and lab test results) and general pretraining data gathered from BooksCorpus and Wikipedia. Table 2 presents the augmented number of data records after operating BioBERT and ClinicalBERT based augmentation and applying various voting techniques of unlabeled data records. The numbers show the counts for each category in the set formed by the combination of the currently labeled data and the augmented labeled data.

Table 2. Class distribution of augmented medical text dataset.

Index	Label	Support (Count)
1	digestive system diseases	7962
2	cardiovascular diseases	1544
3	neoplasms	2429
4	nervous system diseases	6026
5	general pathological conditions	10919

3.4. Data Augmentation via BERT-based Learning

The training dataset, sourced from Kaggle [1], is labeled, whereas the test dataset in this setup is unlabeled. Consequently, the training dataset can only be utilized for training and preliminary testing during the initial phase of evaluation. The labeled dataset is split into 70% for training and 30% for testing, in line with the original configuration provided by the dataset's creator to maintain consistency with baseline results.
The training process has been conducted using the dataset distribution described in Table 1 and the augmented dataset distribution presented in Table 2. The whole flow of the methodology is shown in Figure 1.
The objective of the proposed method starts with evaluation using the initial dataset provided and evaluated in [1], employing ML algorithms such as K-Nearest Neighborhood (KNN), Decision Tree (DT), Random Forest (RF), Stochastic Gradient Descent (SGD) and Support Vector Machine (SVM). Subsequently, the augmented portion of the unlabeled dataset from [1], labeled separately using BioBERT and ClinicalBERT, has been used for classification utilizing the same ML algorithms. Finally, hard and soft voting techniques are applied to the BioBERT and ClinicalBERT labeled data to enhance label accuracy.

4. RESULTS

All experiments were performed on a Google Colab platform equipped with an NVIDIA T4 GPU, which provided sufficient computational resources for training the models. The results of the model trained on the raw dataset are summarized in Table 3. Among the models, the RF model achieved the best performance, whereas the KNN model yielded the poorest results. As detailed in Section 3, we employed BioBERT and ClinicalBERT embeddings with the labeled data available to us. Tables 4 and 5 illustrate that the application of BioBERT and ClinicalBERT embeddings led to improvements in all models.
Tables 6 and 7 present the results of incorporating hard voting and soft voting mechanisms into the BioBERT and ClinicalBERT embeddings, respectively. For hard voting, we employed a majority voting approach, while for soft voting, we utilized a weighted voting strategy. The performance of soft voting has proven to be superior compared to hard voting. Soft voting involves considering the probabilities associated with each class rather than merely the

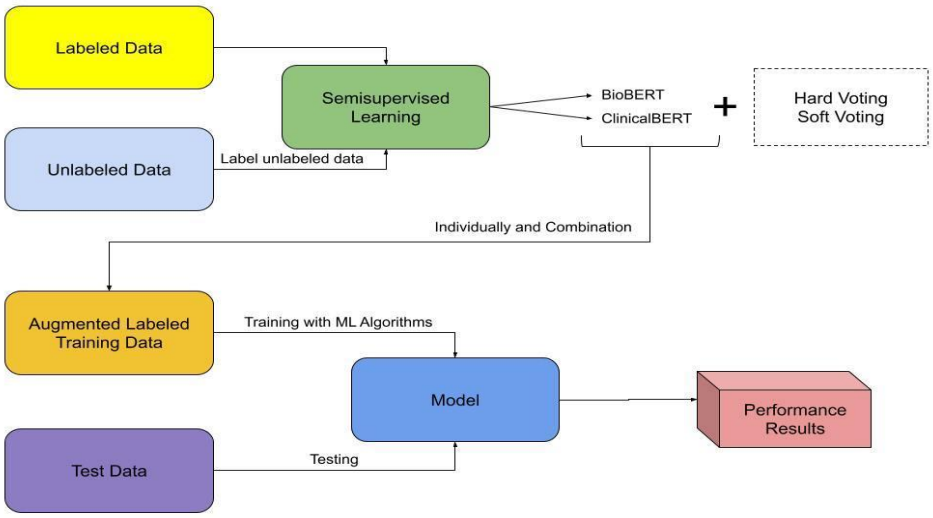

Figure 1. Our proposed method.

predicted class labels. By taking into account the likelihoods, soft voting offers a more comprehensive approach to decision-making. This probabilistic perspective allows soft voting to leverage the uncertainty inherent in the predictions of individual models, thereby enhancing overall predictive performance. Consequently, the results demonstrate that incorporating class probabilities allows soft voting to achieve greater accuracy and robustness in classification tasks.

Table 3. Initial comparison of ML algorithms.

ML Model	Precision	Recall	F1-Score
KNN	0.70	0.68	0.68
Decision Tree	0.82	0.78	0.78
Random Forest	0.79	0.79	0.79
SGD	0.77	0.78	0.77
SVM	0.78	0.78	0.78

Table 4. BioBERT.

ML Model	Precision	Recall	F1-Score
KNN	0.70	0.68	0.68
Decision Tree	0.82	0.78	0.78
Random Forest	0.79	0.79	0.79
SGD	0.77	0.78	0.77
SVM	0.78	0.78	0.78

Table 5. ClinicalBERT.

ML Model	Precision	Recall	F1-Score
KNN	0.75	0.74	0.74
Decision Tree	0.89	0.87	0.87
Random Forest	0.89	0.89	0.89
SGD	0.83	0.84	0.83
SVM	0.86	0.86	0.86

The results in Table 3 show that the RF model achieved the highest F1-score of 0.79, whereas the KNN model yielded the lowest with an F1-score of 0.68. While models like SVM and SGD performed decently, they still didn't perform as well as RF. This initial comparison highlights the varying effectiveness of traditional machine learning models when trained on raw datasets without any domain-specific embeddings.

However, when domain-specific embeddings such as BioBERT and ClinicalBERT were introduced, as shown in Tables 4 and 5, there was a significant improvement in the performance across all models. For instance, the F1-score of the KNN model improved from 0.68 (Table 3) to 0.76 and 0.74 in the BioBERT and ClinicalBERT embeddings, respectively. This increase indicates that incorporating domain knowledge via embeddings helps even weaker models like KNN capture more relevant features from the dataset, thereby enhancing performance.

The most substantial improvement can be seen in the DT and RF models, where the F1-scores jumped from 0.78 and 0.79 (Table 3) to 0.88 and 0.89, respectively, when BioBERT was applied (Table 4). Similar trends are observed with ClinicalBERT, where the F1-score for these models remains at 0.87 and 0.89 (Table 5), further validating that both embeddings significantly benefit models that rely heavily on structured decision-making processes.

Table 6. Hard voting.

ML Model	Precision	Recall	F1-Score
KNN	0.75	0.75	0.74
Decision Tree	0.87	0.85	0.85
Random Forest	0.86	0.86	0.86
SGD	0.83	0.84	0.83
SVM	0.84	0.84	0.84

Table 7. Soft voting.

ML Model	Precision	Recall	F1-Score
KNN	0.77	0.77	0.76
Decision Tree	0.89	0.87	0.87
Random Forest	**0.90**	**0.90**	**0.90**
SGD	0.85	0.85	0.84
SVM	0.87	0.87	0.86

Moreover, the results in Tables 6 and 7 demonstrate the added value of voting mechanisms. Hard voting (Table 6), while improving the performance of models like KNN and Decision Tree, did not outperform the individual best-performing models. For example, the RF model with hard voting achieved an F1-score of 0.86, which is slightly lower than its performance with BioBERT and ClinicalBERT embeddings. The reason for this could be that hard voting does not effectively utilize the probabilities associated with each model's predictions. Instead, it simply chooses the most common class label among the models, which may not capture the nuances of the predictions, especially in complex datasets. As a result, hard voting might not leverage the strengths of individual models to their fullest extent.

However, soft voting (Table 7) yielded the best overall results, particularly for the RF, which achieved the highest F1-score of 0.90. This improvement illustrates the effectiveness of using probability-based ensemble methods, as soft voting allows the model to consider the confidence levels of predictions across different classes. The increase in performance, especially for models like RF and SVM, demonstrates that soft voting not only enhances robustness but also enables the model to make more accurate decisions by leveraging the uncertainties of individual classifiers.

BioBERT and ClinicalBERT embeddings significantly boost the performance of individual models, incorporating soft voting further enhances classification outcomes, making it the most effective approach among the strategies tested. The results suggest that combining domain-specific embeddings with ensemble methods provides a robust solution for improving predictive performance in medical classification tasks.

Table 8 provides a summary of the hyperparameters for BioBERT and ClinicalBERT models. The hyperparameters include the optimizer, learning rate, decay, batch size, number of epochs, max sequence length, number of classification layer, number of neurons in classification layer, activation. Table 9 provides a summary of the hyperparameters used for the Random Forest model that achieved the best F1-score. The hyperparameters include the n_estimators, criterion, min_samples_split, min_samples_leaf.

Table 8. Hyperparameter settings of BioBERT and ClinicalBERT.

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1e-5
Decay	1e-6
Batch Size	16
Number of Epochs	25
Max Sequence Length	384
Number of Classification Layer	1
Number of Neurons in Classification Layer	512
Activation	ReLU

Table 9. Hyperparameter settings of Random Forest.

Hyperparameter	Value
n_estimators	100
criterion	gini
min_samples_split	2
min_samples_leaf	1

5. DISCUSSION

Table 10 provides a comparative summary of key studies in the field of medical text classification, highlighting the architectures, augmentation techniques, voting mechanisms, datasets, target variables, and performance measures. The table illustrates a diverse range of methodologies applied to various datasets, demonstrating the evolution of medical text classification. Studies employed a variety of neural network architectures, ranging from

preliminary models like RNN [26], CNN [27] and LSTM [28] to transformer-based models such as BioBERT, ClinicalBERT, GPT-2, and GPT-3.5 in more recent works [29-31]. Transformer-based models, particularly BioBERT and its variants, have shown superior performance across different tasks, indicating their effectiveness in understanding medical texts.

Earlier studies did not leverage augmentation, relying solely on original datasets. While later studies increasingly utilized synthetic text generation techniques, particularly with models like GPT-2 and GPT-3.5, these efforts primarily focused on generic applications of augmentation [29-31]. In our study, we took a distinct approach by combining BioBERT and ClinicalBERT with carefully designed augmentation techniques, specifically tailored for medical text classification. This novel integration yielded superior results, highlighting the effectiveness of our methodology and improving model performance for specialized medical datasets.

Among the reviewed studies, only one explicitly employed a voting mechanism, highlighting its rarity and significance in medical text classification tasks. In addition to this, our study also utilizes a carefully designed ensemble voting mechanism, further demonstrating its potential to improve performance and address the complexities of medical datasets. This distinguishes our methodology by demonstrating how voting can complement augmentation and transformer-based models, yielding robust and reliable results, particularly in addressing the complexities of medical datasets. This dual focus on augmentation and voting highlights the unique contributions of our study compared to others.

For data augmentation, more advanced and modern models such as GPT-4 can be utilized to generate high-quality synthetic text tailored to the domain. Additionally, implementing adaptive ensemble voting could enhance performance by assigning greater weight to models that excel in specific classes or tasks. Furthermore, incorporating other domain-specific transformer models, such as BiomedBERT and PubMedBERT, alongside BioBERT and ClinicalBERT, can diversify the ensemble and improve overall classification accuracy across a wider range of medical text classification tasks.

While the methodology in our study demonstrates promising results, successful deployment would require addressing key factors such as ensuring data privacy and security, optimizing computational efficiency for scalability, and adapting the models to the specific needs of different institutions. Additionally, providing explainable predictions to build trust among clinicians and integrating the system seamlessly into existing EHR workflows are critical for ensuring reliability and compliance with medical and technical standards. These challenges are generally applicable to all studies involving medical text classification, as they share common issues related to handling sensitive data, managing computational resources, and ensuring practical integration into clinical environments.

Table 10. Comparison of studies on medical text classification.

Study	Architecture	Augmentation	Voting Mechanism	#Documents Considered	Target Variable	Performance Measures
[26]	RNN	-	-	263,706 patients, ~54.61 visits per patient	Medication code	Recall: 85.53%
[27]	CNN	-	-	300,000 patients	Unplanned readmission	AUC: 0.819
[28]	LSTM	-	-	7,191 patients, 53,208 admissions	Unplanned readmission	F1-Score: 0.79
[29]	GPT-2, BioBERT	+	-	55,404 discharge summaries (MIMIC-III dataset)	Unplanned readmission	AUC 0.669, F1-Score 0.3115
[23]	GPT-2, ClinicalBERT	+	-	48,393 generated documents	Unplanned readmission	AUC: 0.822
[24]	BERT, BioBERT, PubMedBERT	+	+	Not specified	Document classification	F1-Score: 0.908
[25]	BioBERT	+	-	3,140 admissions	Medical specialty classification	Micro-F1: 0.882
[30]	GPT-2, BioBERT, BiomedBERT, ClinicalBERT	+	-	73,671 cerebrovascular disease reports	Intracerebral hemorrhage classification	F1-Score: 0.81
[31]	GPT-3.5, RoBERTa	+	-	3,219 radiology reports	Radiology chapter classification	Micro-F1: 0.8846
Our Study	BioBERT, ClincalBERT, Random Forest	+	+	14438 initial records, 28.880 augmented records	Medical specialty classification	F1-Score: 0.90

6. CONCLUSION

In this study, we tackled the challenges of medical text classification by employing semi-supervised learning and BERT-based models to overcome the limitations of labeled data scarcity. Our results demonstrated that integrating domain-specific embeddings such as BioBERT and ClinicalBERT with ensemble methods like soft voting significantly enhances classification performance. For instance, soft voting achieved the highest F1-score of 0.90 with the Random Forest model, underscoring the effectiveness of leveraging probabilistic decision-making in classification tasks.

The results showed a clear enhancement in performance when domain knowledge was integrated through embeddings. For instance, the incorporation of these models allowed even weaker classifiers like KNN to identify more relevant features, improving their overall effectiveness. Additionally, our analysis of different voting mechanisms revealed that while hard voting improved performance for some models like KNN and DT, it did not surpass the best individual model performances. This may be attributed to hard voting's reliance on the most common class labels, which might overlook valuable information encoded in the probabilities of model predictions. In contrast, soft voting proved to be a more effective approach, leading to superior accuracy, particularly for models such as RF. This underscores the importance of leveraging the probabilistic nature of model outputs, which enhances decision-making and ultimately leads to better classification outcomes.

Our results indicate that the combination of data augmentation and contextual BERT-based models effectively addresses the challenges of limited labeled data in medical text classification. We have significantly improved the performance and accuracy of medical text classification tasks by incorporating voting strategies, paving the way for more effective applications in the healthcare domain. Additionally, the models are optimized for the specific vocabulary of clinical contexts, which highlights their potential for domain-specific applications. While the scope is centered on leveraging a specific medical dataset from Kaggle, the methodology and findings provide a strong foundation for further exploration and adaptation to other datasets and domains. Future work could extend these approaches to broader datasets and diverse medical settings.

Author Contributions

The authors contributed equally to this work.

Conflict of Interest

The authors declare that they have no conflict of interest.

REFERENCES

[1] Kaggle, "Medical Text Classification Dataset." Available: https://www.kaggle.com/code/chaitanyakck/medical-text-classification/, (Accessed: Jan. 24, 2025).

[2] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," ACM Computing Surveys, vol. 55, no. 7, pp. 1-39, 2022.

[3] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234-1240, 2020.

[4] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," arXiv preprint, arXiv:1904.05342, 2019.

[5] K. M. Chaitrashree, T. N. Sneha, S. R. Tanushree, G. R. Usha, and T. C. Pramod, "Unstructured medical text classification using machine learning and deep learning approaches," in 2021 IEEE International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), pp. 429-433, 2021.

[6] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," BMC Medical Research Methodology, vol. 22, no. 181, 2022.

[7] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 13, no. 2, pp. 1-41, 2022.

[8] K. Taha, P. D. Yoo, C. Yeun, D. Homouz, and A. Taha, "A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights," Computer Science Review, vol. 54, no. 100664, 2024.

[9] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," Cambridge, MA: MIT Press, 2016.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, 2015.

[11] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corredo, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," Nature Medicine, vol. 25, no. 1, pp. 24-29, 2019.

[12] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE Journal of Biomedical and

Health Informatics, vol. 22, no. 5, pp. 1589-1604, 2017.

[13] L. Torrey and J. Shavlik, "Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques," Hershey, PA: IGI Global, 2010.

[14] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," Journal of Big Data, vol. 3, pp. 1-40, 2016.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 5998–6008, Long Beach, CA, USA, 4–9 December 2017.

[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint, arXiv:1810.04805, 2018.

[17] U. Naseem, K. Musial, P. Eklund, and M. Prasad, "Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding," in 2020 IEEE International joint conference on neural networks (IJCNN), pp. 1-8, 2020.

[18] I. Alimova and E. Tutubalina, "Multiple features for clinical relation extraction: A machine learning approach," Journal of Biomedical Informatics, vol. 103, no. 103382, 2020.

[19] B. Bhasuran, "BioBERT and similar approaches for relation extraction," in Biomedical Text Mining, pp. 221-235, New York, NY: Springer US, 2022.

[20] J. V. A. de Souza, E. T. R. Schneider, J. O. Cezar, L. E. Silva, Y. B. Gumiel, E. C. Paraiso, D. Teodoro, and C. M. C. M. Barra, "A multilabel approach to Portuguese clinical named entity recognition," Journal of Health Informatics, vol. 12, pp. 366-372, 2020.

[21] K. Zeng, Z. Pan, Y. Xu, and Y. Qu, "An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: Algorithm development and validation," JMIR Medical Informatics, vol. 8, no. 7, e17832, 2020.

[22] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," Journal of Big Data, vol. 8, no. 101, 2021.

[23] Q. Lu, D. Dou, and T. H. Nguyen, "Textual data augmentation for patient outcomes prediction," in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2817-2821, 2021.

[24] A. Erdengasileng, Q. Han, T. Zhao, S. Tian, X. Sui, K. Li, and J. Zhang, "Pre-trained models, data augmentation, and ensemble learning for biomedical information extraction and document classification," Database, pp. 1-6, 2022.

[25] H. Zhang, D. Zhu, H. Tan, M. Shafiq, and Z. Gu, "Medical specialty classification based on semiadversarial data augmentation," Computational Intelligence and Neuroscience, Article ID 4919371, 2023.

[26] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in 1st Machine Learning for Healthcare Conference, vol. 56 of Proceedings of Machine Learning Research, PMLR, pp. 301–318, 2016.

[27] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: A convolutional net for medical records," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 22–30, 2017.

[28] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," Journal of Biomedical Informatics, vol. 69, pp. 218–229, 2017.

[29] A. Amin-Nejad, J. Ive, and S. Velupillai, "Exploring transformer text generation for medical dataset augmentation," in Twelfth Language Resources and Evaluation Conference (LREC), Marseille, France, May 2020, pp. 4699–4708, 2020.

[30] Y. H. Kim, C. Kim, and Y. S. Kim, "Language model-based text augmentation system for cerebrovascular disease-related medical reports," Applied Sciences, vol. 14, no. 19, Article ID 8652, 2024.

[31] J. Collado-Montañez, M. T. Martín-Valdivia, and E. Martínez-Cámara, "Data augmentation based on large language models for radiological report classification," Knowledge-Based Systems, vol. 308, Article ID 112745, 2025.