Classification of Medical Specialties through Clinical Text Analysis using Natural Language Processing

Ekaterina Sedykh
University of Tartu
ekaterina.sedykh@ut.ee

Glib Manaiev
University of Tartu
glib.manaiev@ut.ee

Dmytro Fedorenko
University of Tartu
dmytro.fedorenko@ut.ee

Abstract

This project aims to develop a model that can accurately classify medical specialties based on clinical text transcriptions. By leveraging natural language processing techniques, particularly the DistilBERT model, and optimizing it using the Particle Swarm Optimization (PSO) algorithm, we aim to distinguish between various specialties. The developed model can be beneficial for healthcare professionals, medical institutions, and researchers in organizing and managing clinical data, as well as for the development of medical expert systems [1].

1 Introduction

Medical text classification is an important task in the field of healthcare, as it can facilitate the organization and management of clinical data, contribute to the development of medical expert systems, and support research in various medical specialties. With the rapid growth of medical data in the form of electronic health records (EHRs) and clinical transcriptions, there is a need for efficient and accurate methods for classifying and analyzing this data.

In this project, we aim to develop a model that can accurately classify medical specialties based on clinical text transcriptions. We leverage natural language processing (NLP) techniques and pre-trained transformer models, such as BERT and its variants, to identify and extract relevant features from the medical transcriptions, enabling the model to distinguish between various specialties. Our goal is to achieve a high level of classification accuracy and to demonstrate the effectiveness of our approach compared to existing methods in the literature.

2 Related work

The related work section outlines various studies and approaches in the fields of natural language processing and medical text classification. These works include the introduction of the BERT model, which has set new standards in NLP tasks, and its lighter version, DistilBERT. The relevant works explore the application of BERT, a pre-trained transformer model, in the context of clinical text

BERT, introduced by Devlin et al. (2018), revolutionized natural language processing with its ability to generate contextualized word embeddings through deep bidirectional transformers. It is pre-trained on unlabeled text data using tasks like masked language modeling and next sentence prediction. Fine-tuning BERT on specific NLP tasks yields state-of-the-art results. Building on this, DistilBERT was introduced by Sanh et al. (2019) as a distilled version of BERT that provides the benefits of a smaller model with near-equal performance. Leveraging knowledge distillation in the pre-training phase, DistilBERT is 40% smaller, 60% faster, and retains approximately 97% of the language understanding capabilities of BERT. This makes it more suitable for on-device computations and settings with constrained computational resources, without sacrificing the quality of the results.

Li et al. (2023) conducted a comparative study of pre-trained language models for long clinical text. They observed that transformer-based models, known for their exceptional performance, faced challenges with memory and time consumption due to the self-attention mechanism. To address this, they introduced Clinical-Longformer and Clinical-BigBird, two pre-trained models capable of handling larger input sizes. These models outperformed shorter sequence transformers, including ClinicalBERT Huang et al. (2019).

Gasmi (2022) explored the use of BERT for medical text classification tasks. They focused on leveraging pre-trained models for transfer learning, specifically in the context of automatic disease prediction. The authors demonstrated the promising

[1] https://github.com/katesedykh/NLP-medical-specialities

results achieved by fine-tuning BERT on a specific medical dataset. In comparison to other models such as CNN, LSTM, and traditional machine learning approaches, BERT consistently outperformed them.

Rasmy et al. (2021) addressed the challenges of developing deep learning-based predictive models using electronic health records (EHRs). Limited training data availability in EHRs hinders model performance. To overcome this, the authors proposed Med-BERT, a contextualized embedding model pre-trained on a large-scale structured EHR dataset. Fine-tuning experiments showed that Med-BERT significantly improved prediction accuracy, especially for tasks with small training sets. The study highlighted the potential of Med-BERT to enhance disease prediction studies, reduce data collection expenses, and accelerate advancements in AI-aided healthcare.

These papers collectively demonstrate the power of BERT and its variants in the clinical text domain, highlighting their ability to handle long texts, improve classification tasks, and enhance predictive models in healthcare settings.

3 Methods

3.1 Data

The dataset used in this project consists of clinical text transcriptions representing various medical specialties. Each transcription is labeled with the corresponding medical specialty. The dataset used in this project is the Medical Transcriptions dataset, which contains nearly 5000 medical transcription samples across various medical specialties. The number of samples for each specialty ranges from a few dozen to several hundred. The dataset includes clinical text transcriptions and corresponding labels indicating the medical specialty. The dataset is divided into training, validation, and testing sets, with a 70/15/15 split.

3.2 Preprocessing

Since BERT models benefit from the context provided by case information, punctuation, and stop words, our preprocessing steps are limited to:

- Tokenizing the text using the DistilBERT tokenizer.
- Padding and truncating the sequences to a fixed length.
- Creating attention masks to identify padded tokens.

3.3 Model Fine-Tuning

We use the DistilBERT model from the HuggingFace transformers library for our project. DistilBERT is a lighter version of the BERT model, designed for faster training and inference without significant loss in performance. We fine-tune the DistilBERT model on our training set, adapting it to the specific task of classifying medical specialties based on clinical text transcriptions.

3.4 Optimization Algorithm

In our study, the DistilBERT model was optimized utilizing the Particle Swarm Optimization (PSO) algorithm, a population-based optimization technique modeled on the social behavior of birds and fish. We applied the PSO approach as detailed by Gasmi (2022), using this method to pinpoint the optimal parameters for our model.

The PSO algorithm operates by simulating the behavioral pattern of individuals in a flock or school as they gravitate towards their perceived personal best and the overall best position within a multi-dimensional search space. In this context, the multi-dimensional search space signifies the possible hyperparameters for the DistilBERT model, and each 'particle' in the swarm represents a specific configuration of these hyperparameters.

Each particle is assigned a randomized velocity and is navigated through the search space based on its own best solution (pbest) and the global best solution (gbest) identified by the swarm. The position and velocity of each particle are iteratively updated until the optimal solution is discovered. It is crucial to note that the PSO algorithm seeks to minimize -1 times the accuracy of the model, which is why the reported accuracy appears negative.

The implementation of the PSO algorithm was achieved using the PyTorch library, which allowed for smooth integration with the DistilBERT model. For future research, fine-tuning the parameters that govern the PSO velocity, such as the inertia weight 'w' and the cognitive and social scaling parameters 'c1' and 'c2', respectively, could result in improved stability during the optimization process and potentially enhance the performance of the DistilBERT model.

3.5 Evaluation Metrics

To assess the performance of our model, we use evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's performance in classifying medical specialties. Additionally, we visualize the model's performance using a confusion matrix, which helps identify potential areas for improvement.

4 Results

4.1 Data Preprocessing and Label Distribution Analysis

The Medical Transcriptions dataset underwent a preprocessing step in which NaN values were removed. Following this, the distribution of labels was examined, revealing a heavily imbalanced dataset with a few categories having significantly more samples than others. Oversampling techniques were deemed inefficient, and thus a weighted loss approach was initially attempted but yielded unsatisfactory results. To address this issue, the dataset was truncated to include only labels with transcription counts higher than 100 and only 280 samples from the Surgery category were kept, resulting in a more balanced dataset. Figure 1 shows the label distribution of the initial data, data with NaN values removed, and truncated data.

4.2 Model Training and Validation

First, we evaluated the DistilBERT model with the original weights and obtained the baseline results. The results showed that the model had poor performance, with an accuracy of only 0.0585 on the test set (Table 1). The precision, recall, and F1 score were also very low. These results indicate that the model was not usable in its original state.

Next, the PSO algorithm was used to search for the optimal hyperparameters for the model. We obtained good results with the PSO algorithm, with an accuracy of 0.4585 on the test set (Table 2). The precision, recall, and F1 score were also improved significantly.

However, we observed that the validation accuracy curve of the PSO algorithm was unstable, as seen in Figure 2, where it rapidly jumped between good and bad sets of parameters. This instability can be attributed to the parameters responsible for the PSO velocity, which should be reduced in future studies to increase the algorithm's stability.

4.3 Confusion Matrix Analysis

To gain a deeper understanding of the model's performance across different categories, we analyzed confusion matrices for both the validation and test sets (Figure 3). Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class.

Our analysis showed that certain specialities, such as 'Cardiovascular / Pulmonary', 'Consult - History and Phy.', and 'Orthopedic', were predicted accurately by our model. These categories showed a high count on the diagonal of the confusion matrix, indicating a high number of correct predictions.

However, there were also a significant number of misclassifications. These misclassified instances are shown as high counts off the diagonal in the confusion matrix. These results suggest potential areas for improvement in our model, potentially through further tuning of hyperparameters or the use of different model architectures or approaches.

5 Discussion

5.1 Model Performance

Our proposed model, which combines the DistilBERT architecture with the PSO optimization algorithm, demonstrated promising performance in classifying medical specialties based on clinical text transcriptions. The model's accuracy, precision, recall, and F1-score on the test set indicate that it can distinguish between some medical specialties.

However, the model's performance varied across different medical specialties. Some specialties, such as 'Cardiovascular / Pulmonary' and 'Orthopedic', were classified with higher accuracy, while others, such as 'Neurology' and 'Radiology', showed lower performance. This discrepancy may be attributed to the inherent differences in the language and terminology used in each specialty or the varying sample sizes in the training dataset.

We conducted a detailed analysis of transcriptions across different categories and examined the misclassified cases to identify possible reasons for the incorrect predictions made by our model. Upon investigation, we observed that our model frequently confused Surgery with Orthopedics. We discovered that a significant number of misclassified Surgery transcriptions contained mentions of specific body parts, which is a characteristic feature of Orthopedics transcriptions as well. Additionally,

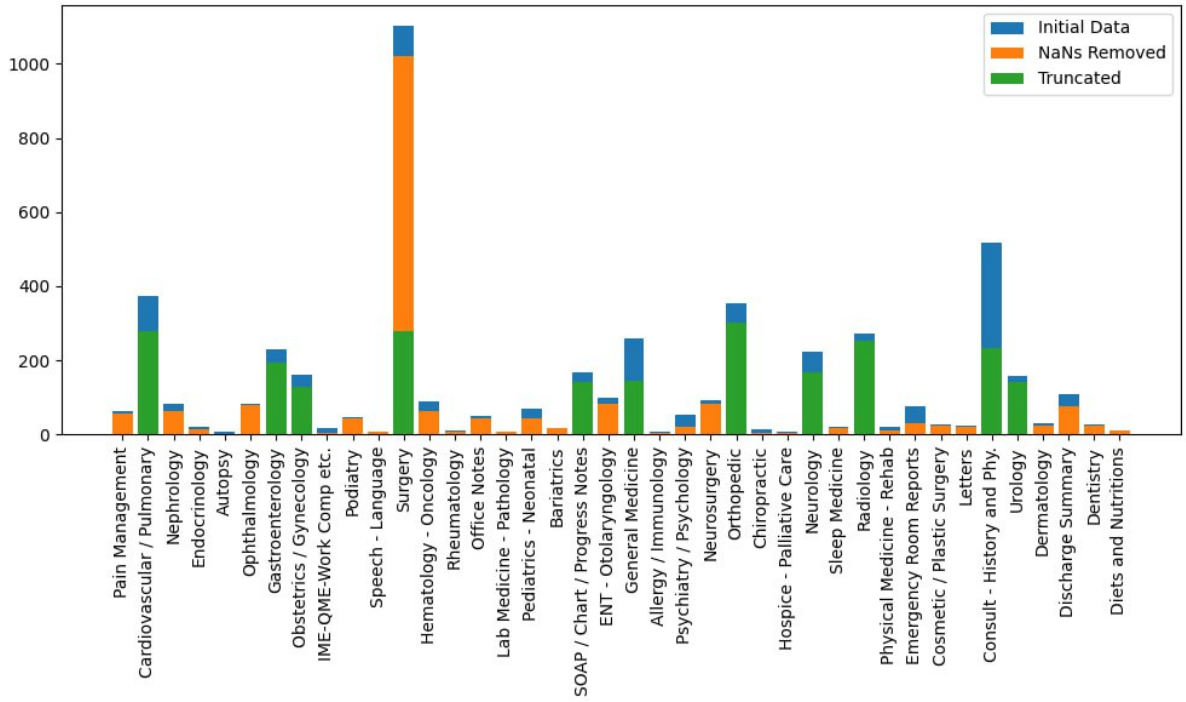

Figure 1: **Label distribution across the Medical Transcriptions dataset.** Bar chart depicting the distribution of labels across the Medical Transcriptions dataset at various stages of data cleaning. The 'Initial Data' bars represent the label distribution in the original dataset. The 'NaNs Removed' bars indicate the distribution after removing instances with missing values, while the 'Truncated' bars show the distribution after truncating less frequent labels. The chart highlights changes in label frequencies at each stage, with some categories like 'Surgery', 'Orthopedic', and 'Cardiovascular / Pulmonary' retaining high counts throughout the data processing stages.

Dataset	Accuracy	Precision	Recall	F1 Score
Eval Set	0.076	0.018	0.076	0.028
Test Set	0.059	0.007	0.059	0.012

Table 1: Results of the baseline model.

Dataset	Accuracy	Precision	Recall	F1 Score
Eval set	0.5126	0.5164	0.5126	0.4746
Test set	0.4585	0.4259	0.4585	0.4201

Table 2: Results of the training with PSO.

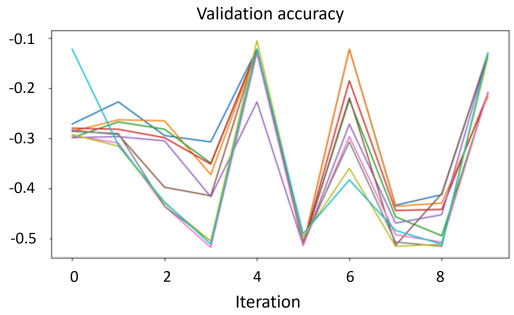

Figure 2: **Validation Accuracy.** Fluctuation of the validation accuracy in the Particle Swarm Optimization (PSO) algorithm. Note that PSO minimizes -1 times the accuracy, which is why the values appear negative. The rapid jumps between high and low accuracy indicate instability in the algorithm's performance, attributed to the parameters controlling PSO velocity. Future studies should aim to reduce this velocity to enhance the stability of the PSO algorithm.

a considerable portion of Urology transcriptions described pre and post-operative diagnoses, potentially leading to confusion with the Surgery category. On the other hand, Consultation, General Medicine, and SOAP transcriptions often included patient-specific details such as age, gender, and other general information, which were less prevalent in other categories. This disparity in content could have contributed to the misclassification of these categories.

This analysis provides valuable insights into the specific areas where our model tends to struggle and misclassify certain categories. By leveraging this knowledge, we can work towards addressing the weaknesses and enhancing the model's performance. Furthermore, this analysis empowers us to make informed decisions about selecting the most suitable model for specific categories, thereby optimizing the overall accuracy and effectiveness of our system.

5.2 Impact of Preprocessing and Data Balancing

After removing the NaN values and balancing the dataset through truncation, the dataset size was significantly reduced. However, this approach led to an improvement in the model's performance. Before balancing, the model was mainly predicting the dominant classes, while ignoring the underrepresented ones. Although the numerical performance was not bad, the model's accuracy was not reliable due to the lack of attention to the underrepresented classes. Therefore, the truncation step was essential to ensure that the model could make accurate predictions across all classes.

5.3 Limitations and Future Work

Despite the promising results, our study has some limitations. First, the model's performance could be further improved by incorporating domain-specific knowledge, such as medical ontologies or structured information from electronic health records, to enhance its understanding of the clinical text. This could be achieved by exploring knowledge-guided models, as discussed in the related work section.

Second, the choice of the PSO algorithm for optimization was really effective, but other optimization techniques may yield better results. Future studies could explore alternative optimization algorithms to further enhance the model's performance. Additionaly we can look into different data preprocessing since DistilBERT without PSO showed low results.

Lastly, our model was trained and evaluated on a single dataset of medical transcriptions. To ensure the model's generalizability, it would be beneficial to test its performance on additional datasets, covering a wider range of medical specialties and clinical scenarios. This would also provide insights

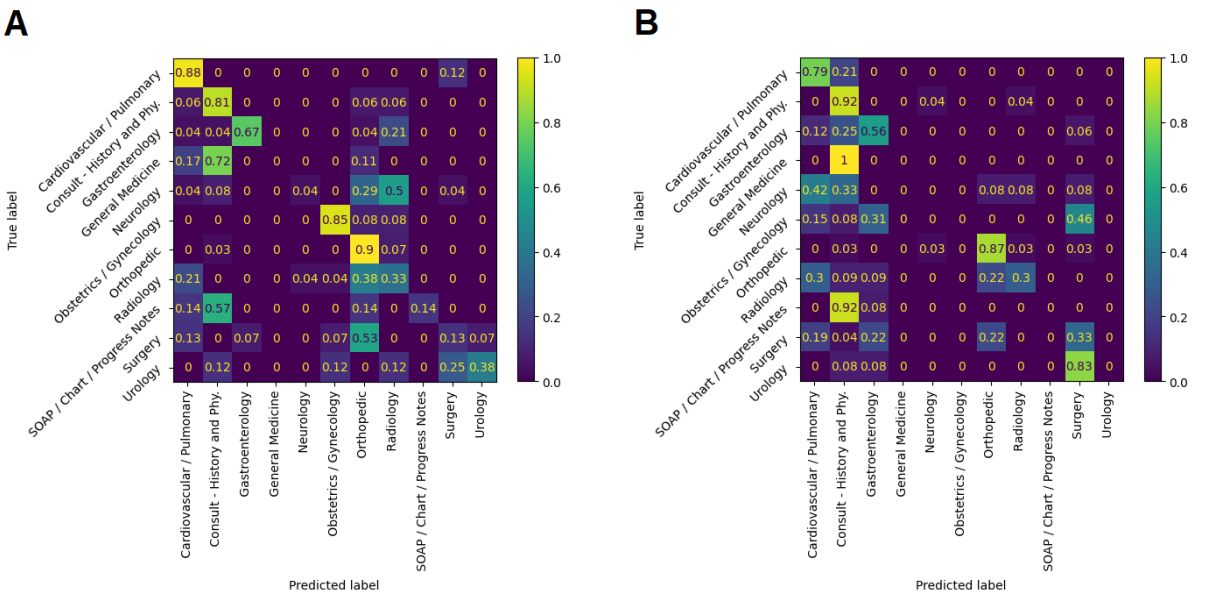

Figure 3: **Confusion Matrix for the Validation (A) and Test (B) Sets.** This matrix illustrates the normalized performance of our model on the validation set. Each entry (i,j) in the matrix represents the proportion of instances of class i that were predicted as class j. The diagonal entries still represent correctly classified instances, while off-diagonal entries now represent the misclassifications as a proportion of the total instances in each class.

into the model's adaptability to other medical text classification tasks or its applicability in other domains.

6 Conclusions

Our study highlighted the potential of combining a pre-trained DistilBERT architecture and Particle Swarm Optimization (PSO) algorithm for the classification of medical specialties based on clinical text transcriptions. Despite the initial imbalance and in our dataset, strategic preprocessing and truncation techniques enabled a significant improvement in the model's performance.

The baseline model's performance, with an accuracy of only 0.059 on the test set, was significantly enhanced by the introduction of PSO, yielding an accuracy of 0.4585. While this demonstrates the benefits of incorporating optimization techniques, it also suggests the room for further enhancements. Variations in performance across different medical specialties indicate that there are opportunities for improving the model's ability to accurately predict underrepresented classes.

The study's findings underline the importance of domain-specific knowledge and data preprocessing in building models for clinical text. It also showed the potential limitations of a single, general-purpose model to accurately capture the nuances across various medical specialties, thereby highlighting the necessity for model fine-tuning, exploration of alternative optimization techniques, and potentially domain-specific models.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Karim Gasmi. 2022. Improving bert-based model for medical text classification with an optimization algorithm. In *Advances in Computational Collective Intelligence: 14th International Conference, ICCCI 2022, Hammamet, Tunisia, September 28–30, 2022, Proceedings*, pages 101–111. Springer.

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

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.