National College of Ireland

RoBERTa-Based NLP System for Enhanced Disease Prediction from Symptom Descriptions

MSc Research Project
MSc Data Analytics

Karthikeya Anusury
Student ID: 22217096

School of Computing
National College of Ireland

Supervisor: Abid Yaqoob

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Karthikeya Anusury…….

Student ID: ……22217096…….

Programme: ……MSc Data Analytics…… **Year:** ……2023-2024……

Module: ……MSc Research Project……

Supervisor: ……Abid Yaqoob……

Submission Due Date: ……16-09-2024……

Project Title: RoBERTa-Based NLP System for Enhanced Disease Prediction from Symptom Descriptions

Word Count: ……6730…… **Page Count**……20……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Karthikeya Anusury……

Date: ……16-09-2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

RoBERTa-Based NLP System for Enhanced Disease Prediction from Symptom Descriptions

Karthikeya Anusury
Student ID: 22217096

Abstract

Due to the fast-changing nature of medical treatment, there is an urgent need for rapid translation with precision into employable medical interpretations to conduct timely diagnoses of the disease. In this context, the work has designed and compared the performance of state-of-the-art deep learning natural language processing models BERT and RoBERTa and machine learning models XGBoost and Random Forest for disease prediction from symptom descriptions. In this study, the SymptomsDisease246k dataset was applied, and those diseases that possess at least 1000 samples were selected for reliability and statistical significance.

All the models were tested for their handling of medical terminology, computational efficiency, and predictive accuracy. The methodology involves rigorous preprocessing of data, model implementation, and then an extensive analysis based on model performance metrics like accuracy, F1 score, precision, and recall.

Among these three transformer-based models, significant improvement in accuracy is visible as compared to the performance of traditional machine learning. BERT slightly outperformed RoBERTa during short-term training, having the best accuracy of 0.8712 and F1 score of 0.8720. During longer training, this model was more stable than both, though BERT still had a slight edge. XGBoost turned out to be a strong baseline, providing the possibility to balance performance and computational efficiency, yielding 0.8578 accuracy. Random Forest, while being less accurate, was the fastest in training.

The current study has demonstrated the role that transformer-based models can play in the field of medical text analysis, but at the same time, it has also shown that traditional methods of machine learning can be helpful in certain contexts. The results suggested that when choosing a model to use in healthcare, there is a trade-off between performance and efficiency. Moreover, in a clinical environment, this needs human expert validation. This study further adds to the development of AI-assisted clinical decision-making processes with widened insight into both the strengths and limitations of various modeling 5approaches in medical text analysis.

1 Introduction

The accurate reading of the medical texts and the clinical description of symptoms become only more important than ever for timely diagnosis when the pace of change in the environment of healthcare is getting faster and faster. With increasing medical information, the need for robust and reliable NLP tools to help professionals in this task also rises (Devlin et al., 2019; Liu et al., 2019). Of these, the leading models in NLP domains that have proved highly able in many language comprehension tasks are BERT and RoBERTa. However, how they stand against each other in performance within this highly specialized field of medical text analysis, especially in disease prediction from symptom descriptions, has relatively remained underexplored.

The most obvious gain will be to the health sector. More accurate and faster processing of medical texts may be realized, yielding speedier diagnoses, reduced error rates, and more efficient deployment of medical resources. Not only such but in academic circles, a deeper understanding of how these models perform with medical terminology can also inform the development of even more specialized NLP tools for healthcare applications.

Although both BERT and RoBERTa have been very successful with general tasks associated with the use of language, the often complex and detailed nature of medical terms poses a challenge. The texts are normally full of domain-specific jargon, abbreviations, and syntactic structures not commonly occurring in everyday language use. This brings into question what adaptability and performance of these models this specialty domain can receive.

Previous studies have proven the power of BERT and RoBERTa separately for medical applications. For example, Peng et al (2019). showed that BERT performed very well in clinical concept extraction, while Alsentzer et al (2019). applied it in clinical natural language inference. Meanwhile, Lewis et al (2020). showed the potential of RoBERTa in biomedical question answering. An important lacuna in the literature seems to be that no direct comparison between these models concerning the specific task of disease prediction from symptom descriptions has been drawn.

This research will seek to fill this gap and further compare BERT and RoBERTa more rigorously in the use of medical text analysis. Therefore, this work has the following objectives:

- To build and fine-tune the BERT and RoBERTa models upon a large data set of symptom-disease pairs.
- To assess and compare the performance of both models with metrics that are based on measures such as accuracy, F1 score, precision, and recall.
- Determine if identified models can handle the usage of medical jargon and complex forms of symptom description.
- To test how effective the models are and how much computational power is needed for them in this specific task.
- To assess the implications of the findings for potential clinical applications.

The success of the above objectives will be measured through comprehensive statistical analysis of the performance of different models, detailed error analysis, and critical evaluation of the results in a real-world clinical setting.

Methodologically, this paper is based on the quantitative approach using the SymptomsDisease246k[1] dataset, available from Hugging Face. that would include data preprocessing, model implementation with aid provided by library Transformers, tightly regulated training and evaluation procedures, and detailed comparative analysis. The focus on diseases represented by (1000+) samples gives statistical significance and covers the most frequent medical conditions.

This research contributes to the scientific literature by providing:

- A complete head-to-head comparison between BERT and RoBERTa in medical text analysis. And implementation of basic ML models for additional comparison of these transformer models.

[1] https://huggingface.co/datasets/fhai50032/SymptomsDisease246k

- Insights into the strengths and weaknesses of these models in the handling of specialized medical terminology.
- An overview of the implications of these findings that are practical for clinical applications

The structure of this report is as follows: section 2 provides a detailed review of related work, situating this study within the existing literature; Section 3 outlines the methodology of the research, including data preparation, model implementation, and evaluation procedures. Section 4 gives design specifications of the implemented models. Section 5 details the implementation. The outcome of the experiments is presented in Section 6, together with a detailed analysis and discussion. Finally, Section 7 concludes the report with essential findings, limitations, and proposed directions for future research.

2 Related Work

The application of advanced NLP models in medical text analysis has seen remarkable development. The present paper critically surveys the current state of research into BERT and RoBERTa models within medical contexts, that is, evaluation of their strengths, limitations, and potential applications.

2.1 Advancements in NLP Models for Medical Text Analysis

Recent studies have shown the effectiveness of BERT and RoBERTa models in processing complex sets of medical data. Abdal et al. (2023) demonstrated that RoBERTa is efficient in sentiment analysis on small, informal messages; therefore, it has implications for the processing of medical notes with fragmented structures. This study proves the flexibility of RoBERTa but does not touch directly on medical terminology, hence leaving a wide scope for investigation within healthcare contexts.

Built on this with the identification of Personal Health Mentions in social media content using Roberta Khan et al. (2022). This work demonstrated RoBERTa's potential in health-related information extraction from unstructured text but was based on patient narratives—a key facility needed during processing. The focus on social media content might not be representative of the full complexities of professional medical documentation and hence be limiting in any direct applicability to a clinical setting.

In a more domain-specific setting, Pal et al (2021). probed the performance of RoBERTa in natural language questions-to-SQL queries. This work reflects more of the power of RoBERTa on complex queries apart from the medical database, portraying its accuracy and efficiency in tasks that have a deep understanding of the language. The latter work is focused on SQL conversion, relevant but not directly related to symptom-to-disease mapping at the heart of medical diagnosis.

2.2 Domain-Specific Adaptations and Applications

Developed a variant fine-tuned on biomedical applications, BioMed-RoBERTa Monea and Marginean (2021. The latter achieved better results on natural language inference (NLI) and recognition of question entailment (RQE) tasks applied to medical domain-specific texts. Although this paper proves that in-domain fine-tuning is possible, it does not regard comparing

the result of BioMed-RoBERTa with other models like BERT over medical domains, which leaves room for comparative analysis.

Showed that speech analysis could be used with RoBERTa for the prediction of Alzheimer's disease, which thus gave huge potential for this model in diagnosing diseases by their verbal symptoms Wang et al (2023). Although it is very relevant research to medical diagnostics oriented by speech analysis and not on written descriptions of symptoms, it gives room for exploration in diagnosis by text.

2.3 Comparative Studies and Model Enhancements

Comparative studies on BERT and RoBERTa in medical contexts are scant, though some research gives an idea of their relative performance. Zhao et al (2021). proposed a BERT-based model for sentiment analysis and key entity detection in financial texts. While this has nothing to do with medical applications, the work showcases BERT's prowess in the extraction of crucial information from complex texts and provides a sort of template that could be used in comparison with traditional approaches to machine learning.

Further enhanced RoBERTa by equipping it with an underline facility for co-attention mechanisms; they resulted in better machine reading comprehension Kim et al. (2023). This will allow the attention to be directed selectively at the relative part of the difficult texts, mainly very useful for the analysis of many extensively medical detailed documents. The fact that it showed better metrics of performance from the study implies deriving possible gains, without a baseline comparison in medical contexts regarding BERT, within the analysis of medical text.

2.4 Emerging Trends and Future Directions

Recent trends in NLP research are oriented more toward specialized applications of BERT and RoBERTa in healthcare. For instance, Kamatha et al (2022). worked on emotion detection using EmoRoBERTa and presented the potential of this model to understand patient emotions, which is a core component of psychological evaluations. This proves the adaptability of RoBERTa-based models but does not take into consideration the specific challenges of disease prediction from symptom descriptions.

Another rising trend is the incorporation of domain-specific knowledge into the pre-trained models. Lewis et al. (2020) developed BioBERT a biomedical language representation model pre-trained on large-scale biomedical corpora, which has shown some bright prospects in bridging the gap between general-purpose language models and specialized medical applications. However, comparative studies between BioBERT and other models like RoBERTa in medical text analysis are still very few.

2.5 Research Gap and Study Rationale

The existing literature review has shed light on a few gaps in research studies undertaken so far:

- Limited direct comparison: Although individual studies have shown the efficacy of BERT and RoBERTa in classifying medical texts or disease prediction from symptom descriptions, there has been a less wide and fair comparison between medical text analysis and these models.

- Insufficient focus on medical specificity: Most studies showcase the capabilities of the models based on general language tasks, which mostly fall under specific medical subtasks; very few rise to face the challenge of precisely interpreting complex medical terminology and the relationships of symptoms with diseases.
- Lack of large-scale evaluation of symptom-disease mapping: Most of what has been done is task and medical condition-specific. This leaves a void in the rich understanding of how these models perform across diseases and symptoms.
- Lack of performance analysis in clinical diagnostic contexts: While some studies briefly touch on medical applications, there is an insufficient body of research regarding how performance generalizes to real-world clinical diagnosis scenarios.

These gaps in the literature justify the need for a comprehensive, comparative study of BERT and RoBERTa concerning disease prediction from symptom descriptions. The present study will attempt to bridge these gaps by providing valuable insights into the relative strengths and limitations that these models have in medical text analysis, which may turn out beneficial to be considered in developing further improvements in AI-aided clinical diagnostics.

3 Research Methodology

This section explains the methodology that will be followed in this research work: preparation of data, model implementation, training procedures, and evaluation techniques as shown in Figure 1. In a view to achieve the objectives of the study, which was to compare BERT and RoBERTa models for medical text analysis with relation to disease prediction from symptom descriptions, a proper research procedure must be designed.

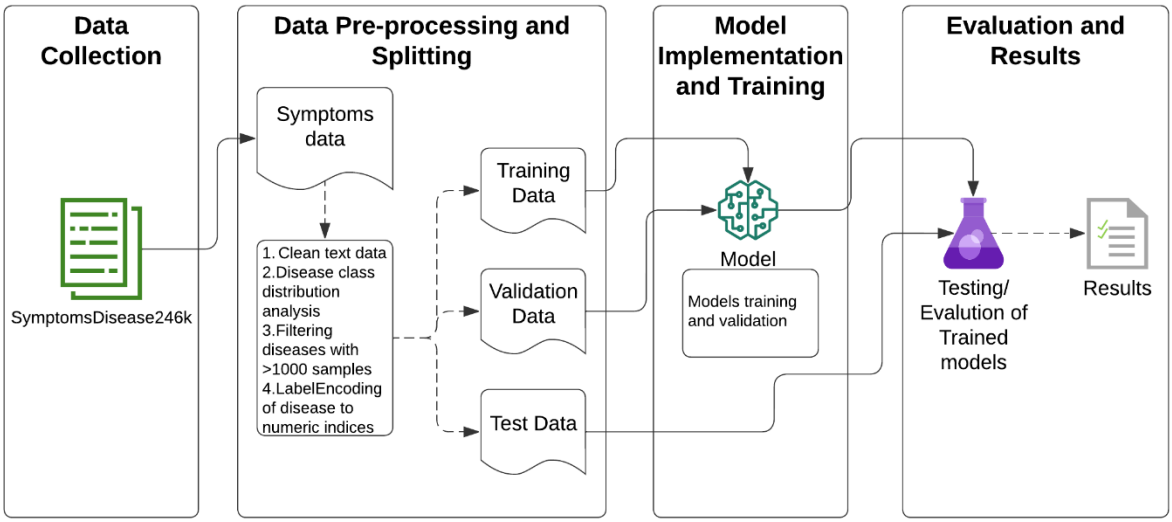

Figure 1 Research Methodology

3.1 Computational Environment

The research environment used Google Colab as the primary computational platform, powered by its A100 GPU with 40GB of memory (Carneiro et al., 2018). This created a very high-performance environment within which large-scale medical text data could be processed efficiently and complex transformer models trained expeditiously.

3.2 Dataset and Preprocessing

3.2.1 Data Source

This study was based on the SymptomsDisease246k dataset, sourced from Hugging Face. The choice of this dataset for research is due to the rich aggregation within the symptom-disease pairs and for how very conducive it would be in training models related to medical NLP and evaluating their performance.

3.2.2 Data Preprocessing

The preprocessing phase involved several key steps:

- Load the dataset from a Hugging Face dataset.
- Clean the text data by lowercasing and stripping it of unwanted punctuation marks.
- Class distribution analysis of diseases that are represented with an adequate number as shown in Figure 2.
- Initially, there were 1,546 unique diseases in the dataset out of which the top 20 diseases are shown in Figure 3 with the highest sample size. After merging similar classes (e.g., 'you may have flu' and 'flu' were considered as one class), the number was reduced to 773 unique diseases. After filtering for diseases with at least 1000 samples, the final number of unique diseases was 201.
- Encode the labels with sci-kit-learn's LabelEncoder, mapping the names of the diseases to numeric indices (Hancock & Khoshgoftaar, 2020).

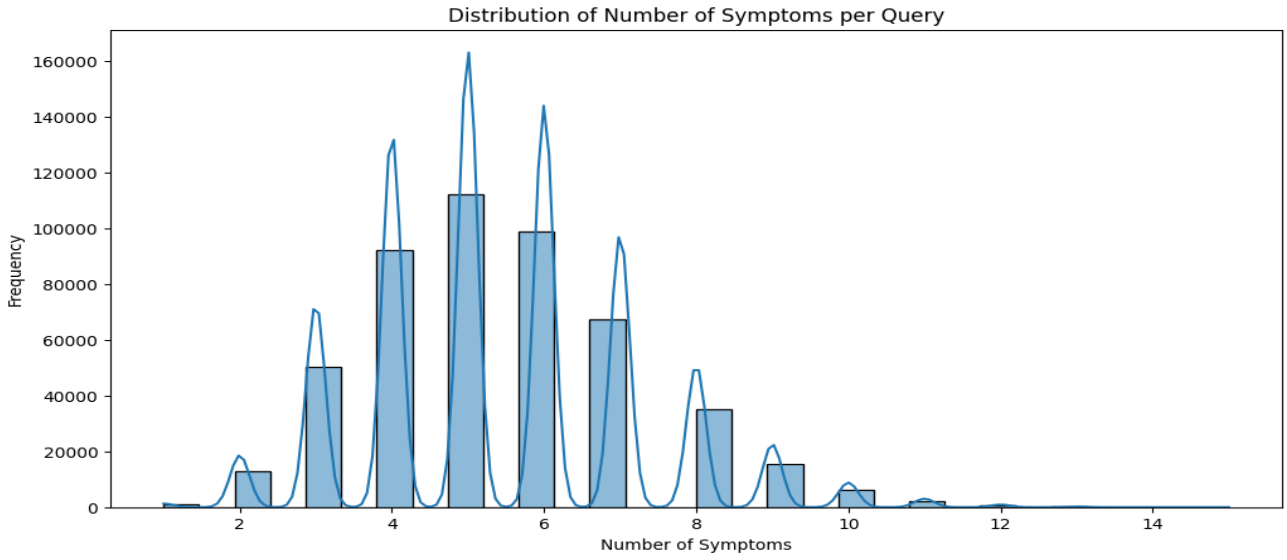

Figure 2 Distribution of Number of Symptoms per Query

3.2.3 Data Splitting

The pre-processed dataset is stratified and split into training (72%), validation (8%), and test sets (20%). This makes sure that the stratification reaches good representative disease distribution in all sets.

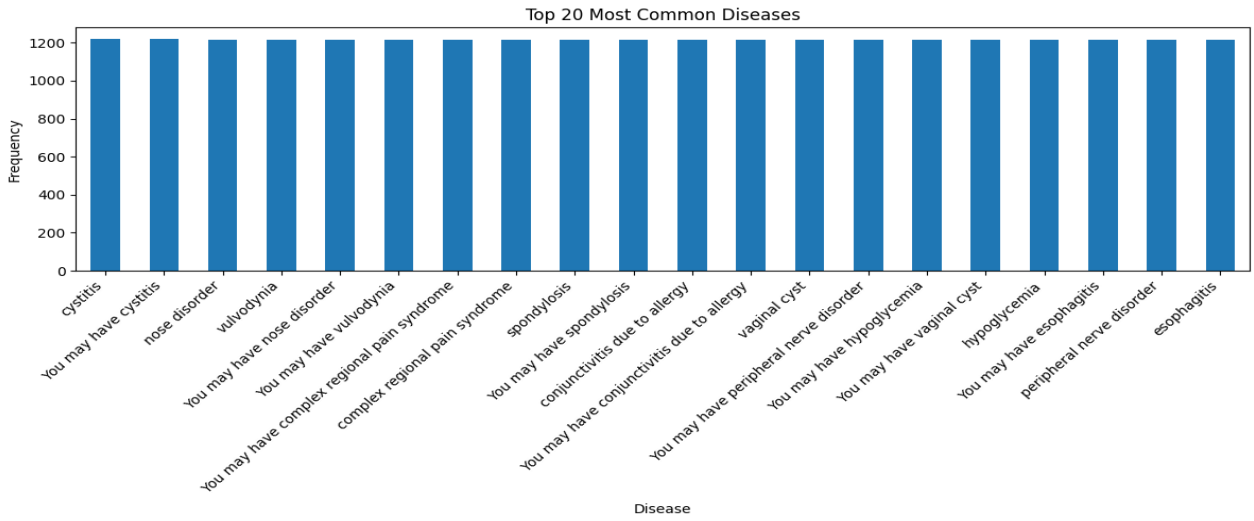

Figure 3 shows the Top 20 most common diseases with large samples

3.3 Transformer Models Implementation

Two transformer-based models were implemented:

1. BERT: Using the pre-trained 'bert-base-uncased' model from Hugging Face[2].
2. RoBERTa: Loading 'Roberta-base' pre-trained model from Hugging Face[3].

Both transformer-based models were adapted for sequence classification with the number of output labels equivalent to the count of unique diseases in the filtered dataset (201 classes after filtering for diseases with at least 1000 samples).

3.4 Tokenization

Since the tokenization approach is similar for both BERT and RoBERTa, this will subsequently include examples that will illustrate the tokenization process for examples.

Table 1: Comparison of Tokenization in BERT and RoBERTa

Feature	**BERT Tokenization**	**RoBERTa Tokenization**
Tokenizer	BERT Tokenizer	RoBERTa Tokenizer
Tokenization	WordPiece (subwords with '##')	Byte-level BPE (word boundaries with 'Ġ')
Special Tokens	[CLS]' (start), '[SEP]' (end)	<s>' (start), '</s>' (end)
Example: "Symptoms include hypertension and tachycardia	['[CLS]', 'symptoms', include', 'hyper','##tension','and', 'tach', '##ycard', '##ia','[SEP]']	[<s>, 'Symptoms', Ġinclude, Ġhyper, tension, Ġand, Ġtach, ycard, ia, '</s>]
Sequence Length	128 (truncation/padding)	128 (truncation/padding)

3.5 Training Procedure

3.5.1 Hyperparameters

[2] https://huggingface.co/docs/transformers/model_doc/bert
[3] https://huggingface.co/docs/transformers/model_doc/roberta

The training process for both BERT and RoBERTa models utilizes a set of hyperparameters: A learning rate of 2e-5 was considered; a low learning rate prevents drastic changes in the pre-trained. A batch size of 128 was used for its tradeoff between computational efficiency. The optimizer used here is AdamW which is more appropriate for transformer models like these, with a high number of parameters (Hassan et al., 2022). A linear schedule with a warmup was adopted, which helps stabilize early training and allows updates with finer steps as training proceeds (Wu & Liu, 2022).

3.5.2 Training Loop

The training process was structured to efficiently update the model parameters while updating performance. It processes a batch of 128 samples for each iteration, performing a forward pass through the model to compute the loss using cross-entropy (Mao et al., 2023). The training process then uses backpropagation to calculate gradients, which are used by the AdamW optimizer to update the models' parameters to minimize the loss (Lee & Park, 2023). Validation is included in the loop at the end of each epoch to monitor the performance on unseen data. Early stopping with two epochs is implemented to prevent the overfitting problem which halts the training if the validation loss does not improve.

3.5.3 Model Saving

To ensure optimal performance and enable further analysis, two versions of each model architecture were saved:

1. Best Model: The model state with the lowest validation loss during training was saved. This typically represents the model's state with the best generalization capability.
2. Final Model: The model state corresponding to the best overall performance throughout training was saved, along with the model state at the end of training, whether due to early stopping or reaching the maximum number of epochs.

3.5.4 Training

Besides initial runs with 5 and 10 epochs, tests up to 30 epochs were conducted to investigate long-term learning dynamics, identify potential overfitting, determine when performance plateaus, evaluate computational efficiency, and assess model stability for practical applications.

The results of the longer training runs were analyzed with particular care and compared to the shorter training runs so that a complete view was possible concerning each model's behavior and performance characteristics for several different training runs.

3.6 Evaluation Methodology

The evaluation of BERT, RoBERTa, and baseline ML models employed a comprehensive approach to thoroughly assess their performance in disease prediction from symptom descriptions. The best-performing models were used to generate predictions on a held-out test set. Various evaluation metrics, including accuracy, F1 score, precision, and recall, along with a confusion matrix, were computed. Misclassifications were analyzed to identify patterns and areas for potential improvement. This detailed evaluation methodology provides a fine-grained understanding of the strengths and weaknesses of each model.

Comparing performances using multiple metrics and conducting in-depth analysis establishes a strong basis for comparing BERT and RoBERTa in this medical text classification task.

3.7 Comparative Analysis

To facilitate a direct comparison between BERT and RoBERTa:

- Performance metrics for both models were compiled and compared.
- Training time and computational efficiency were analyzed.
- In-depth error analysis for both models has been done to understand the strengths and weaknesses associated with medical terminology and their symptom descriptions.

3.8 Baseline Machine Learning Models Implementation

In addition to transformer-based models (BERT and RoBERTa), two classic machine learning baselines were implemented:

- XGBoost (eXtreme Gradient Boosting): XGBoost is t fast-optimized distributed gradient boosting library designed to be highly effective, flexible, and portable. It is known for its performance in multiple machine learning competitions and is right suited for large-scale data. Key configurations: n_estimators: 100, learning_rate: 0.1, max_depth: 5, random_state: 42 and n_jobs: -1 (utilize all available cores)
- Random Forest: Random Forest is an ensemble learning technique; it constructs lots of decision trees during the training process. It is chosen due to its robustness, but also because it can work on high-dimensional data with the potential of not reaching overfitting. Key configurations: n_estimators: 10, random_state: 42 and n_jobs: -1 (utilize all available cores)

Most of these models have been implemented and evaluated using the same rigorous methodology as the transformer models, to make sure there is a fair comparison between all approaches.

4 Design Specification

This section details the architectural design, and technical specifications of the system implemented for the comparison of BERT and RoBERTa models in medical text analysis. It goes to some extent into the details of the framework underneath, along with various design decisions that back up the methodology described in the previous section. It contains four main cardinal components that work together in harmony for the integration of the model.

4.1 Data Processing Module

The Data Processing Module forms the basis of this system, which undertakes most of the prime activities related to loading, cleaning, and preparation of data in a nutshell, optimally formatting the input data for model training.

1. Data Loading: Use the Hugging Face datasets library for fast uploads and large data volumes.
2. Preprocessing Pipeline: This part implements an end-to-end preprocessing pipeline, eventually covering text cleaning and encoding of the labels for each sentence.

3. Stratified Sampling: The mechanism for stratified sampling, to ensure a balanced representation of diseases in the training, validation, and test sets is maintained.

4.2 Model Implementation Module

This module brings a uniform interface for working with BERT and RoBERTa models:

- Model Initialization: This class uses the transformers library to initialize pre-trained models.
- Tokenization Wrapper: This provides model-specific tokenization, BERT, and RoBERTa, in front of a consistent interface to the rest of the system.
- Sequence Classification Adaptation: Modifies the output layer of pre-trained illness classification models to match the number of distinct diseases in the dataset.

Figure 4 compares the workflow of the BERT and RoBERTa showing the primary distinction in the tokenization methods, embedding layer, and pooling strategies.

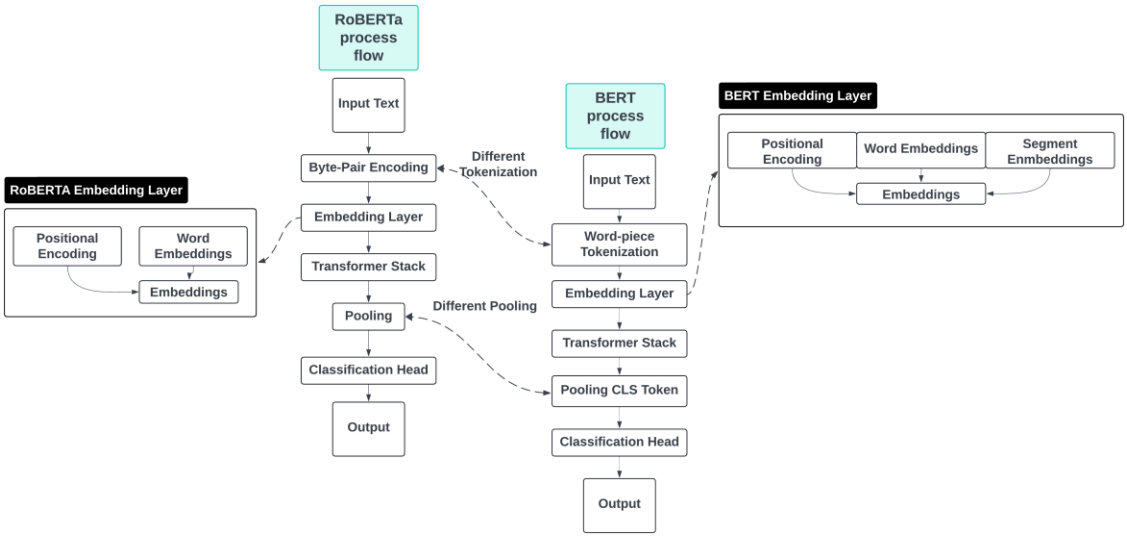

Figure 4 BERT vs RoBERTa Architecture

4.3 Training and Evaluation Module

The Training and Evaluation Module manages the complexities of model training and assessment. The implementation employs a flexible training loop with early stopping efficacy in the training processes. The AdamW optimizer and linear learning rate scheduler with the warmup phase are used for fine-tuning transformer models. Model checkpointing is introduced to ensure the best-performing model based on validation loss is saved. A comprehensive evaluation framework utilizes Scikit-learn's classification report to ensure consistent model assessment.

4.4 Comparative Analysis Module

This module allows for direct comparison between BERT and RoBERTa models.

1. Performance Metrics Compilation: Accumulates and compiles the performance metrics accuracy, F1 score, precision, and recall for both models into a comparable format.

2. Visualization Tools: Uses matplotlib to generate performance curves, so one can draw a pictorial comparison of training dynamics between both models.
3. Results Logging: Implements a structured logging system that saves detailed results and model configurations, facilitating reproducibility and further analysis.

5 Implementation

This section details the last step in the implementation, geared at outputs and tools used in developing a comparison system for BERT and RoBERTa in medical text analysis.

5.1 Transformed Data

This ranges from the most important preparations for the SymptomsDisease246k dataset, getting it ready for model training and evaluation, to data transformation. This involves a more detailed explanation of the transformed data:

1. Pre-processed Dataset: The initial dataset with 1546 distinct disease labels is pre-processed by keeping only diseases with 1000+ samples, resulting in a final dataset of 201 unique disease classes and 336,998 samples.
2. Encoded Labels: Disease names were encoded into numeric indices for machine-learning compatibility by assigning unique integers.
3. Dataset Split: The dataset was then divided into a Training set (for model training), a Validation set (for fine-tuning parameters), and a Test set (for evaluation).

5.2 Developed Models

Two primary transformer models, BERT and RoBERTa, were fine-tuned for disease classification. The BERT and RoBERTa models utilized the pre-trained 'bert-base-uncased', and 'roberta-base' architectures. Both models incorporated a pre-trained transformer base and a classification head with 201 output neurons, aligning with the number of unique diseases identified after filtering the dataset.

For comparison and baseline performance, two traditional machine learning models were also implemented: XGBoost and Random Forest. Both models utilized TF-IDF vectorization with a maximum of 2500 features for text representation and were trained on the same pre-processed dataset as the transformer models, containing 336,998 samples across 201 disease classes.

5.3 Evaluation Outputs

The implementation produced several key evaluation outputs:

1. Performance Metrics: Accuracy, F1 score, precision, and recall for both models.
2. Classification Reports: Detailed per-class and average performance statistics.
3. Training and Validation Curves: Loss and accuracy progression plots during training.
4. Performance and Efficiency Comparison: Direct comparison of BERT and RoBERTa across various metrics and with ML models

5.4 Tools and Languages Used

The implementation utilized:

- Python 3.8.10: Primary programming language.
- PyTorch 1.10.0: Deep learning framework for model building and training.
- Transformers Library (Hugging Face): For accessing pre-trained models and tokenizers.
- Pandas 1.3.4: Data manipulation and preprocessing.
- Scikit-learn 1.0.1: Data splitting, label encoding, and evaluation metrics calculation.
- Matplotlib 3.4.3 and Seaborn 0.11.2: Visualization creation.
- Google Colab: Development environment with A100 GPU.

6 Evaluation

This section contains a detailed result analysis with performance comparisons across BERT and RoBERTa models in the medical text analysis area for disease prediction from symptom descriptions.

6.1 Training Dynamics and Results of BERT and RoBERTa

The initial stage of training is conducted with 5 and 10 epochs and increased to 30 epochs to examine the training and validation accuracies of both models.

6.1.1. Experiment 1 / Training Epoch 5

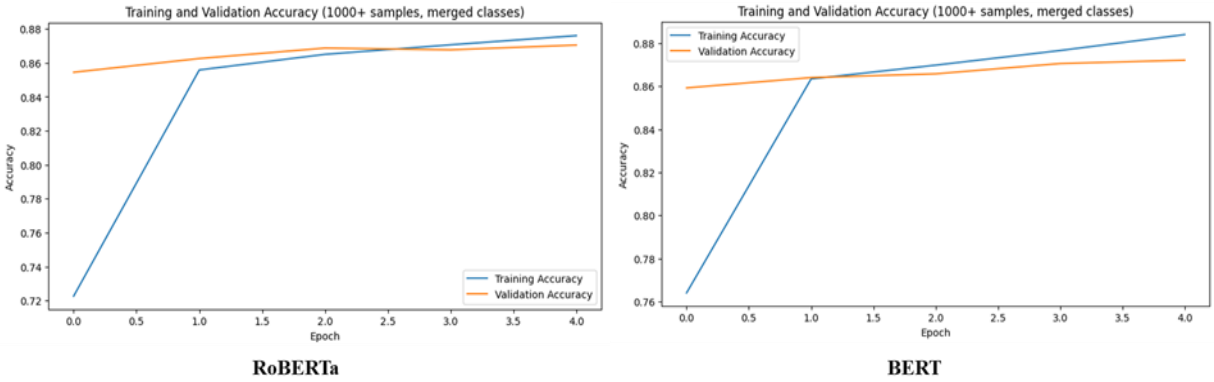

Figure 5 Line graph showing training and validation accuracies for BERT and RoBERTa over 5 epochs

Key observations from experiment 1 (training the models over 5 epochs):

- Both models showed fast improvements in the first epoch, with significant improvements in both training and validation accuracies.
- RoBERTa and BERT models had the early consistency much faster as they achieved a validation accuracy of 0.8545 and 0.8593 respectively in barely an epoch's training.
- BERT is more consistent in improvement throughout all the epochs. RoBERTa significantly decreased in improvement rate after the third epoch.
- In the end, BERT attained a marginally better accuracy in validation (0.8721) compared with RoBERTa (0.8705)

.

6.1.2. Experiment 2 / Training Epoch 10

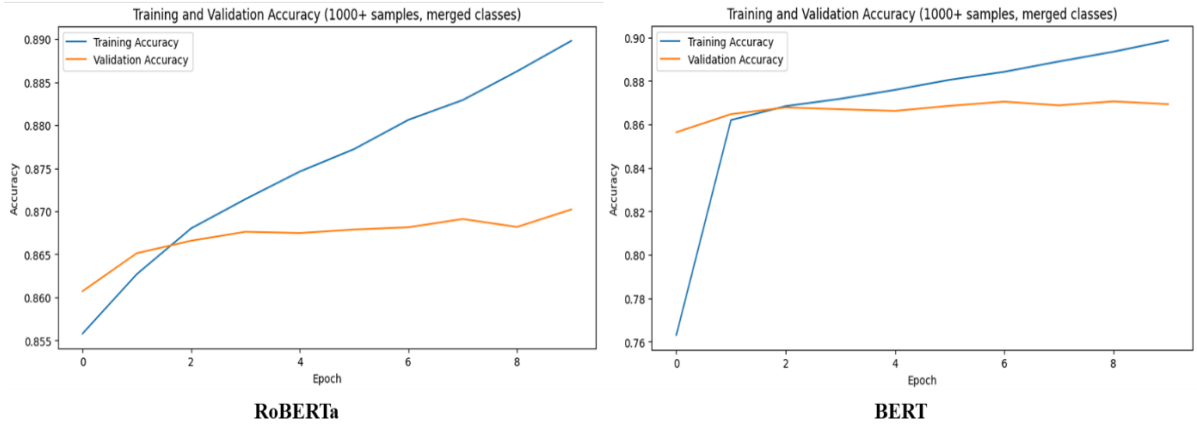

Figure 6 Line graph showing training and validation accuracies for BERT and RoBERTa over 10 epochs

Key observations from experiment 2 (training the models over 10 epochs):

- Both models' training accuracy increased with an increase in several epochs. BERT starts with a lower training accuracy but surpasses from Epoch 4 onward, by Epoch 10 BERT has slightly higher training accuracy (0.8987) than RoBERTa (0.8898).
- RoBERTa maintains a slightly higher validation accuracy than BERT across all the Epochs. However, the validation accuracy of BERT is decreasing at the end of epoch 10.
- BERT shows a potential sign of overfitting, with a wide gap between training (0.8987) and validation accuracy (0.8694) in later epochs, while RoBERTa has a smaller and more consistent gap.

6.1.3. Experiment 3 / Training Epoch 30

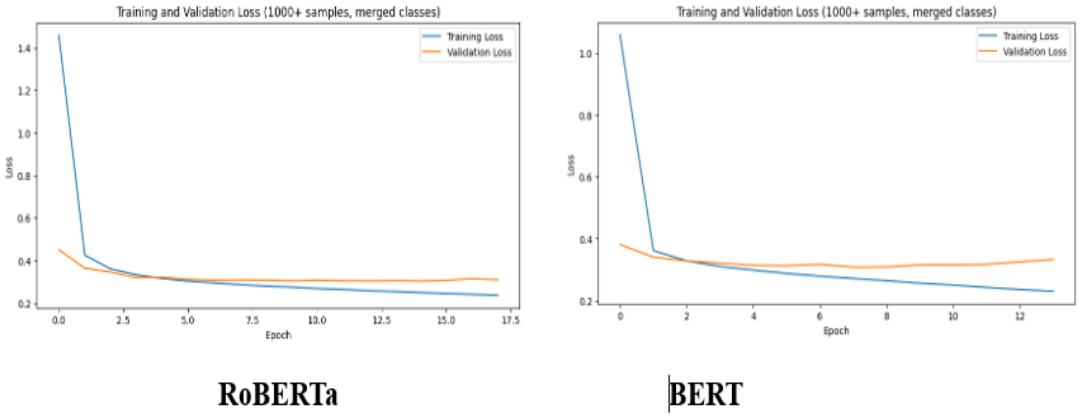

Figure 7 Line graph showing training and validation loss for BERT and RoBERTa over 30 epochs

To further investigate the learning dynamics of the BERT and RoBERTa models, the training was extended up to a maximum of 30 epochs with early stopping. A plot of training and validation loss curves for BERT and RoBERTa over the extended training is presented in Figure 7, respectively.

Key observations from experiment 3 (training the models over 30 epochs with early stopping):

1. BERT Performance: During the additional training, the training loss of BERT continues smoothly down and reaches its minimum at approximately epoch 14. On the other hand, the validation loss reaches a plateau after about 10 epochs and slightly grows during the last epochs. That might indicate that BERT when training on this dataset, could be overfitting if it is trained for too long.
2. RoBERTa Performance: RoBERTa has a similar pattern to BERT but differs in some key points: While the training loss goes down more linearly after the initial rapid drop, its validation loss starts to show an upward trend sooner, at around epoch 6. That perhaps would indicate that RoBERTa is being overfitted for this task a little bit more than BERT.

6.2 Machine Learning Models Results

Complementing the transformer-based models are two traditional machine learning models: Random Forest and XGBoost, namely BERT and RoBERTa. These models have been chosen for the comparison of a baseline and in terms of computational efficiency versus performance.

The results of these ML models are presented in the following table:

Table 2: Machine Learning Models Performance Comparison

Model	Accuracy	F1 Score	Precision	Recall	Training Time (minutes)
XGBoost	0.8578	0.8586	0.8616	0.8578	5.68
Random Forest	0.8244	0.8245	0.8254	0.8244	0.27

Key observations:

1. Performance: XGBoost outperforms Random Forest in all metrics. It yields the maximum value of accuracy, F1 score, precision, and recall.
2. Efficiency: Random Forest is much faster to train; it takes approximately 0.27 minutes, which compares to XGBoost taking 5.68 minutes.
3. XGBoost Performance: XGBoost does very well with an accuracy of 0.8578 and an F1 score of 0.8586-only a bit worse than transformer-based models.
4. Random Forest Performance: The performance is a little worse, yet still decent: Random Forest follows with an accuracy of 0.8244 and an F1 score of 0.8245.
5. Precision and Recall: Both models presented a balanced precision and recall score, which indicates consistent performance across different classes.
6. Computational Efficiency: Both the ML models are significantly faster than the transformer-based models.

6.3 Models Comparative Analysis and Discussion

This section compares performance metrics for all implemented models, namely: BERT, RoBERTa, XGBoost, and Random Forest. The comparison was performed by considering accuracy, F1 score, precision, recall, and training time.

Table 3: Comprehensive Model Performance Comparison

Model	Accuracy	F1 Score	Precision	Recall	Training Time (minutes)	Epochs
BERT (Initial)	0.8712	0.8720	0.8779	0.8703	100.76	5/5
BERT (10 Epochs)	0.8677	0.8678	0.8736	0.8677	201.87	10/10
BERT (Extended)	0.8692	0.8697	0.8776	0.8692	281.98	14/30
RoBERTa (Initial)	0.8691	0.8701	0.8790	0.8692	92.87	5/5
RoBERTa(10 Epochs)	0.8682	0.8686	0.8737	0.8682	185.5	10/10
RoBERTa (Extended)	0.8675	0.8673	0.8730	0.8675	333.68	17/30
XGBoost	0.8578	0.8586	0.8616	0.8578	6.04	N/A
Random Forest	0.8244	0.8245	0.8254	0.8244	0.27	N/A

Analysis and discussion:

- Performance Hierarchy: The BERT and RoBERTa versions perform with a better score of all metrics compared to the XGBoost and Random Forest classical models. Initial training of BERT has the highest accuracy of 0.8712 and enjoys the best F1 score at 0.8720 of all models.
- Extended Training Impact: Extended training (14/30 epochs) for BERT just has a slight drop in accuracy from 0.8712 to 0.8692, while precision and recall remain comparable. RoBERTa too shows extended training with 17/30 epochs, which marginally reduces the performance for all metrics. This puts BERT at an advantage in this setting, considering the marginal superiority and shorter training time in extended runs.
- Traditional ML Models: XGBoost does a great job, achieving an accuracy of 0.8578 and an F1 score of 0.8586. Since these two scores are significant, coming from XGBoost and significantly cutting down the training time of the transformer models, it could be considered a very impressive performance.
- BERT vs. RoBERTa: Results show that BERT has a slight edge over RoBERTa in both the extended training settings. This also puts BERT at an advantage in this setting, considering the marginal superiority and shorter training time in extended runs.
- Role of Traditional ML Models: Despite lower metrics, there is still some valuable insight from XGBoost and Random Forest: a) It gives the baseline performance based on which one can tell how much improvement transformer models provide. b) In cases where computational power is limited or when speedy deployment is required, such models offer an option, especially XGBoost.
- Scalability and Computational Resources: During training, the traditional ML models are much more efficient. The transformer models needed significantly more time for

training. While BERT and RoBERTa clearly show the best results, this performance comes at a cost: increased computational requirements.

Ultimately, though BERT and RoBERTa yield the best performances for disease prediction, it is important to balance model selection and trade-offs. It can be concluded that for transformer models, 5 epochs of training is sufficient. When high accuracy is indispensable, transformer-based models are preferable. For applications needing faster training or deployment, XGBoost strikes a good balance.

6.4 Input validation using BERT and RoBERTa

Along with testing the model with the test dataset, for additional validation custom string inputs are generated with symptoms, and RoBERTa and BERT pre-trained models for over 10 epochs are considered for this validation.

Sample Test Case: A custom input is generated with symptoms of the disease “ingrown toe nail”. This input is fed to the model to predict the disease.

Custom Input: “A patient presenting with abnormal appearing skin, neck swelling, foot or toe pain, and swelling, along with an infected appearance of the skin on the leg or foot, a swollen eye, and irregular appearing nails, may be indicative of a systemic infection or inflammatory condition. Immediate medical evaluation is necessary to determine the underlying cause and initiate appropriate treatment.”

BERT Prediction: ingrown toe nail

RoBERTa Prediction: ingrown toe nail

7 Conclusion and Future Work

This study addressed the research question: "How do BERT and RoBERTa differ in their ability to interpret and analyze medical texts, and what implications do these differences have for their use in clinical diagnosis?" The investigation also compared these transformer models with traditional machine-learning approaches like XGBoost and Random Forest.

BERT outperformed others with the best accuracy of 0.8712 and F1 score of 0.8720 for preliminary training over 5 epochs. Neither BERT nor RoBERTa showed significant gains with longer training, indicating that 5 epochs of training should be enough to learn most of the learnable patterns in the medical text data. Traditional ML models, in particular XGBoost and Random Forest, were much faster when training, while transformer models performed significantly better. This trade-off in terms of efficiency vs. performance thus provides valuable options for different scenarios in medical text analysis.

These transformer models have shown high performance in terms of such complex medical terminologies and symptom descriptions, which will potentially improve the clinical decision-making process. Residual error rates of about 13% underline the importance of human expert validation in a clinical setting.

There are some limitations to this study. Focusing on diseases with more than 1000 samples leads to the loss of possibly rich insights into rarer conditions. Large variants of transformer models and the potential for domain-specific pretraining were not explored. In practice, the choice of model must be made based on specific use-case needs, balancing accuracy requirements with computational constraints and deployment speed. The study emphasizes that AI can advance transformative improvements in the medical analysis of texts, while continuous

research is necessary to address current limitations and ethical considerations in Healthcare AI applications.

Future directions should focus on multimodal learning, integrating diverse medical data for a comprehensive basis for disease prediction. More interpretable models should be developed to enhance their trustworthiness and adaptability in a clinical setting. Longitudinal studies, adapting these models to process time series of patient histories, could provide a clearer understanding of disease progression and long-term care.

Future research should also address the challenge of rare diseases, improving model performance for less frequent conditions. Specialized transformer variants, such as MedBERT, capable of handling extensive medical corpora and improving performance in healthcare settings, need to be developed. Ethical considerations, including possible biases, data privacy, and model equity, must be addressed. Guidelines on responsible deployment in clinical settings should be established.

References

M. N. Abdal, M. H. K. Oshie, M. A. Haue and K. Islam, "A Transformer Based Model for Twitter Sentiment Analysis using RoBERTa," 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ICCIT60459.2023.10441627.

A. A. Khan, F. Kamal, N. Nower, T. Ahmed and T. M. Chowdhury, "An Evaluation of Transformer-Based Models in Personal Health Mention Detection," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 1-6, doi: 10.1109/ICCIT57492.2022.10054937.

D. Pal, H. Sharma and K. Chaudhuri, "Data Agnostic RoBERTa-based Natural Language to SQL Query Generation," 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2021, pp. 1-5, doi: 10.1109/I2CT51068.2021.9417888.

B. Kumar, Sheetal, V. S. Badiger and A. D. Jacintha, "Sentiment Analysis for Products Review based on NLP using Lexicon-Based Approach and Roberta," 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/IITCEE59897.2024.10468039.

P. Yu and Y. Liu, "Roberta-based Encoder-decoder Model for Question Answering System," 2021 International Conference on Intelligent Computing, Automation and Applications (ICAA), Nanjing, China, 2021, pp. 344-349, doi: 10.1109/ICAA53760.2021.00070.

J. -H. Kim, S. -W. Park, J. -Y. Kim, J. Park, S. -H. Jung and C. -B. Sim, "RoBERTa-CoA: RoBERTa-Based Effective Finetuning Method Using Co-Attention," in IEEE Access, vol. 11, pp. 120292-120303, 2023, doi: 10.1109/ACCESS.2023.3328352.

L. Zhao, L. Li, X. Zheng and J. Zhang, "A BERT-based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts," 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Dalian, China, 2021, pp. 1233-1238, doi: 10.1109/CSCWD49262.2021.9437616.

A. M. Monea and A. N. Marginean, "Medical Question Entailment based on Textual Inference and Fine-tuned BioMed-RoBERTa," 2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 2021, pp. 319-326, doi: 10.1109/ICCP53602.2021.9733687.

Y. Wang et al., "Exploiting Prompt Learning with Pre-Trained Language Models for Alzheimer's Disease Detection," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095993.

R. Kamath, A. Ghoshal, S. Eswaran and P. Honnavalli, "An Enhanced Context-based Emotion Detection Model using RoBERTa," 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/CONECCT55679.2022.9865796.

Peng, Y., Yan, S., & Lu, Z. (2019). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In Proceedings of the 18th BioNLP Workshop and Shared Task (pp. 58-65).

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72-78).

Lewis, P., Ott, M., Du, J., & Stoyanov, V. (2020). Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In Proceedings of the 3rd Clinical Natural Language Processing Workshop (pp. 146-157).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1810.04805

Carneiro, T., Medeiros Da Nobrega, R.V., Nepomuceno, T., Bian, G.-B., De Albuquerque, V.H.C. and Filho, P.P.R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, [online] 6, pp.61677–61685. doi:https://doi.org/10.1109/access.2018.2874767.

Hancock, J.T. and Khoshgoftaar, T.M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 7(1). doi:https://doi.org/10.1186/s40537-020-00305-w.

Hassan, E., Shams, M.Y., Hikal, N.A. and Elmougy, S. (2022). The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study. *Multimedia Tools and Applications*. doi:https://doi.org/10.1007/s11042-022-13820-0.

Lee, S. and Park, H. (2023). Effect of Optimization Techniques on Feedback Alignment Learning of Neural Networks. doi:https://doi.org/10.1109/icaiic57133.2023.10067047.

Lu, H., Ehwerhemuepha, L. and Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology*, 22(1). doi:https://doi.org/10.1186/s12874-022-01665-y.

Mao, A., Mohri, M. and Zhong, Y. (2023). *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2304.07288.

Tan, K.L., Lee, C.P. and Lim, K.M. (2023). RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis. *Applied Sciences*, [online] 13(6), p.3915. doi:https://doi.org/10.3390/app13063915.

Wu, Y. and Liu, L. (2022). Selecting and Composing Learning Rate Policies for Deep Neural Networks. *ACM Transactions on Intelligent Systems and Technology*. doi:https://doi.org/10.1145/3570508.