PATIENT NARRATIVES TO SPECIALIST PREDICTION: A BERT-BASED NLP APPROACH FOR AUTOMATIC MEDICAL SPECIALTY CLASSIFICATION

MD. REJWAN RASHID

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This thesis titled on "**Patient Narratives to Specialist Prediction: A BERT-Based NLP Approach for Automatic Medical Specialty Classification**", submitted by **Md.Rejwan Rashid (ID: 221-35-1016)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

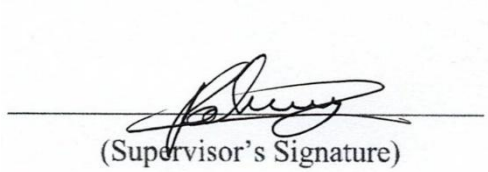
Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.



(Supervisor's Signature)

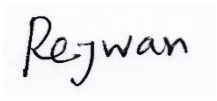
Full Name : Tapushe Rabaya Toma

Position : Assistant Professor

Date : 21.12.25

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.



(Student's Signature)

Full Name : Md.Rejwan Rashid

ID Number : 221-35-1016

Date : 21-12-25

PATIENT NARRATIVES TO SPECIALIST PREDICTION: A BERT-BASED NLP APPROACH FOR AUTOMATIC MEDICAL SPECIALTY CLASSIFICATION

MD.REJWAN RASHID

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Software Engineering)

DAFFODIL INTERNATIONAL UNIVERSITY

NOVEMBER 2025

ACKNOWLEDGEMENTS

On First of all, I want to thank my thesis supervisor Tapushe Rabaya Toma for all the help, advice and support she gave me during this research. The information she shared together with her comments helped establish the direction for this research project. I also want to thank the professors in the Department of Software Engineering at Daffodil International University for their great lectures and for creating an environment of academic excellence.

I am thankful to my friends who supported me with their help and encouragement during my thesis research.

I am grateful to my family because they demonstrate constant love and backing for all my activities. The fact that they believed in me has been what kept me focused on finishing this work. Without all of their help and support, I wouldn't have been able to do this.

ABSTRACT

Healthcare documentation management depends on medical specialty classification to achieve efficient organization and management of medical records. The research investigates the performance of transformer-based deep learning systems when used for medical transcription classification to achieve accurate medical domain identification. The dataset used consists of a large number of real clinical notes labeled with one of 29 medical specialties. The text-based data introduces an original approach to implement NLP technology for healthcare applications. The research creates automated medical specialty identification systems from unprocessed doctor-patient conversation recordings to enhance healthcare operational efficiency and patient referral accuracy.

The research starts by reviewing existing studies about medical text classification and transformer models including DistilBERT. The research includes complete details about the dataset origin and volume and all preprocessing operations that were performed. Special attention is given to handling class imbalance, where underrepresented specialties are assigned appropriate class weights during training. The data preparation process is emphasized to ensure reliability and relevance of the content used in training.

The model evaluation depends on four performance metrics which include top-1 accuracy and top-3 accuracy and macro F1-score and weighted F1-score. The proposed model reached a top-1 accuracy of 54.03% while delivering a top-3 accuracy of 95.61% which proves its capacity to generate multiple suitable predictions for each input. The evaluation process of these metrics takes place at various checkpoints to enable comparison. The model's output results undergo confusion matrix analysis to identify which specialties have the most misclassification errors.

The research demonstrates that transformer-based NLP models achieve success in medical documentation automation and specialty classification work. The paper presents its current boundaries and future research paths which involve using domain-specific models together with ensemble methods. The research investigates actual clinical documentation problems to develop better AI healthcare solutions and build deep learning systems that understand medical texts.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EMR	Electronic Medical Report
NLP	Natural Language Processing
HER	Electronic Health Records
SVM	Support Vector Machine
BoW	Bag of Words
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory network
GRU	Gated Recurrent Unit
CNN	Convolutional Neural Network

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

1.1 Background

The patient starts their healthcare experience by sharing their symptoms and medical background and personal worries through spoken words. Medical staff and doctors use this information to determine the appropriate medical specialty for diagnosis and treatment and for making referrals. Specialty routing and clinical triage function as the system which directs patients to their correct medical treatment. The analysis of clinical text data has experienced a complete transformation because of NLP and transformer models including BERT & DistilBERT (Devlin, 2019). The process of manual interpretation requires extensive time and produces variable results which depend on healthcare provider experience levels particularly when working in rural or telemedicine settings. Clinical medical narratives exist as unstructured content because they follow different writing patterns and lack standardized organization (Meystre, 2008). The availability of electronic medical records (EMR) and transcription datasets and Natural Language Processing (NLP) advancements enables the development of automated systems which analyze clinical texts to recommend specialist care. The transformer-based models BERT and DistilBERT demonstrate the best results in medical language understanding because they excel at processing complex medical terminology and abbreviations and contextual information. Therefore, this thesis aims to develop and evaluate a machine learning model that predicts the correct medical specialty based on a patient's transcription or narrative. The proposed model takes raw text as input and outputs a medical specialty such as Cardiology, Neurology,

Orthopedics etc. The system provides support for both early disease detection and medical referral choices and telemedicine operations.

1.2 Problem Statement

The process of identifying appropriate medical specialties from patient descriptions of their symptoms proves to be both difficult and based on personal judgment because of noise, inconsistency, and domain-specific language. The wrong interpretation of test results may result in delayed medical care and incorrect doctor recommendations and additional healthcare expenses. Medical datasets containing unbalanced information distribution patterns cause models to incorrectly identify rare medical specialties (He, 2009). There is currently no widely used automated system in Bangladesh or many developing regions that can intelligently recommend the appropriate medical specialist from patient-described symptoms.

Therefore, there is a need to develop an NLP-based automated system that can accurately predict the correct medical specialty from clinical transcriptions or patient narratives.

1.3 Research Questions

The research investigates the following set of questions :

- Can the NLP model DistilBERT shows its capability to perform medical transcription classification into specific medical specialties?
- The model achieves what accuracy and F1-score and Top-3 accuracy on real-world medical transcription data?

- Can this automated system assist healthcare platforms, telemedicine and hospital triage by reducing manual referral errors?

1.4 Research Objectives

General Objective:

A system needs to be created which uses Natural Language Processing (NLP) to identify the correct medical specialty from patient notes and recorded speech. The transformer model DistilBERT provides efficient classification task performance through its fine-tuning capabilities (Sanh, 2019).

Specific Objectives:

- The process demands work with actual medical transcription data which needs cleaning and preprocessing.
- To fine-tune a transformer-based model (DistilBERT) for multi-class medical specialty classification.
- The model requires evaluation through Accuracy and Macro-F1 score and Weighted-F1 score and Top-3 accuracy metrics.
- To compare different checkpoints and select the best-performing model.
- To analyze misclassified cases for better understanding and improvement.

1.5 Scope of the Study

- The study investigates 29 different medical fields which include Cardiology, Neurology and others.
- The dataset consists of English-language transcribed data only.
- The study uses transformer-based deep learning models (specifically DistilBERT).
- The system generates the most probable specialty choice but it does not offer medical treatment or healthcare recommendations.
- The model exists within a controlled environment where Python and GPU-based training methods were used for development and testing.
- The application of domain-specific transformers produces better results for biomedical classification tasks (Lee, 2020).

1.6 Limitations

- The dataset contains an imbalance because certain medical specialties appear only a few times while others appear numerous times.
- The model cannot predict rare or unseen specialties not included in the dataset.
- The accuracy of transcription results decreases when transcriptions become both lengthy and noisy because unclear or ambiguous language becomes more difficult to understand.

- The system operates as an instrument which helps doctors during their work activities through referral support and decision-making assistance.

1.7 Thesis Structure

- The research problem along with its objectives and study reasons appear in the first chapter.
- The second part of this dissertation contains a review of current medical NLP research together with existing studies on specialty classification.
- The third chapter contains information about methodology and dataset description as well as preprocessing and model architecture.
- The fourth chapter presents the results of implementation together with evaluation metrics and confusion matrix analysis and evaluation results.
- The final section of the paper includes a summary of findings together with an evaluation of study boundaries and recommendations for additional research.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

Medical text processing encounters major challenges because clinical narratives present disorganized information that contains different types of noise which require medical expertise to understand correctly (Meystre, 2008). The chapter reviews existing research about NLP in healthcare and medical text classification and clinical decision support systems and transformer models BERT and DistilBERT. The research examines existing studies to determine their knowledge gaps and evaluate their results and constraints which shows how this thesis fills the acknowledged research gap.

2.2 Natural Language Processing (NLP) in Healthcare

Artificial Intelligence contains Natural Language Processing as its subfield which enables machines to process and understand human language. NLP allows healthcare systems to automatically extract medical information from electronic health records which enhances their clinical decision-making abilities (Kreimeyer, 2017). The healthcare sector employs NLP for various applications:

- The procedure for extracting data from clinical notes and electronic health records (EHR).
- The system depends on automated processes to generate ICD codes for medical diagnosis needs.

- The system needs to identify diseases based on patient reports for conditions such as diabetes and depression and COVID-19.
- Medical question-answering systems and chatbots.
- Clinical decision support and triage systems.

The healthcare sector faces additional obstacles when implementing NLP because medical documents include abbreviations and fragmented sentences and technical medical language and specialized terminology.

2.3 Traditional Machine Learning Methods

Medical text classification used traditional machine learning models as its main approach until deep learning became popular.

Table 2.1 : Traditional Algorithms for text classification

Algorithm	**Features Used**	**Limitations**
Naive Bayes	Bag-of-Words (BoW)	Poor context understanding
SVM (Support Vector Machine)	TF-IDF vectors	Struggles with long text
Random Forest	Handcrafted features	Requires feature engineering
Logistic Regression	N-grams	Does not capture semantic meaning

The extraction of medical information from unstructured text becomes difficult for Naive Bayes and SVM models because they depend on Bag-of-Words and TF-IDF features which fail to recognize contextual information (Miner, 2020).

2.4 Deep Learning Approaches

Deep learning technology allowed neural networks to achieve superior results than traditional models in their operational capabilities.

Table 2.2 : Deep Learning Models Used in Text Classification

Model	Description	Limitations
CNN (Convolutional Neural Networks)	Detects particular patterns that exist in text data	Fails to perform well when dealing with long sequences
RNN / LSTM	Understands sequential text	Slow training & limited long-term memory
Bi-LSTM / GRU	Better context capturing	Still weaker than transformer-based models

Medical text classification receives improvement from deep learning models which use CNNs and LSTMs to detect patterns in sequential and local data (Hughes, 2017).

2.5 Transformer Based Models

The self-attention mechanism in Transformers enabled models to process all words in a sentence at once which brought a revolutionary change to NLP.

Table 2.3 : Transformer-Based Models

Model	Description
BERT	Pre-trained on large text corpora (Wikipedia + BookCorpus). The system identifies context information through analysis of both left and right directions.
DistilBERT	A lighter and faster version of BERT. 40% fewer parameters, 60% faster, while retaining 95% of BERT's accuracy.
BioBERT, ClinicalBERT	The models trained using biomedical research papers and clinical notes as their source of information.

Because of their superior contextual understanding, BERT-based models are now widely used for medical classification tasks. The medical text classification field has experienced a revolution through BERT and its variants DistilBERT and BioBERT because these models use self-attention to achieve deep contextual understanding which outperforms traditional feature-based models (Lee, 2020).

2.6 Related Work

In recent years, researchers have increasingly applied machine learning (ML) and natural language processing (NLP) techniques to the healthcare domain to automate tasks such as disease prediction, patient record classification, and treatment recommendation. The research indicates that medical text understanding has

transitioned from using Naive Bayes and Decision Trees to BERT and its derivatives because these transformer-based models deliver better results.

Salunke et al. The authors (2015) conducted a research study which compared Naive Bayes and Decision Tree and SVM algorithms for healthcare data classification. The model produced acceptable results but it only worked with organized data and failed to understand free-text patient information (Salunke, 2015).

The 2013 paper "Recommender System Using Collaborative Filtering Algorithm" analyzed how collaborative and content-based filtering methods produce individualized medical suggestions (Alluhaidan, 2013). The system proved successful for user-based suggestions yet it faced two main challenges because of sparse data and insufficient ability to understand medical language in context.

The authors published their work in Sustainability (2023) to study deep learning models CNNs and transformer architectures for healthcare data interpretation. The research demonstrated that transformer-based systems which include BERT achieve superior performance than traditional models when processing unstructured medical data that requires contextual understanding. The systems need powerful computational resources for operation which creates difficulties when trying to use them in limited resource settings (Torres-Ruiz, 2022).

The U.S. patent US10127359 introduced an automated diagnostic support system which combined analytics with NLP for clinical decision-making. The research demonstrated how to merge medical data analysis with text-based interpretation yet

the system operated as a proprietary system which limited its use for research applications (Blue, 2018).

The research depends on the previous studies to develop a DistilBERT-based model that uses BERT for medical specialty classification and recommendation tasks. The system accepts unstructured patient stories as input to identify the most suitable medical department or specialist among Cardiology, Neurology and Dermatology. The model handles class imbalance through weighted loss strategy which maintains equal representation of all 29 specialties.

The model demonstrates high performance through its 54.03% Top-1 accuracy and 95.61% Top-3 accuracy which shows the correct specialty appears in the top three recommendations in nearly every case. The system proves its ability to unite medical text classification with intelligent recommendation through this demonstration which provides a functional and efficient and scalable solution for healthcare systems.

The research unites two independent methods from previous studies which connect clinical knowledge from text-based sources to computerized medical specialty recommendation platforms. The research works reviewed in this paper establish the theoretical and technical basis for this thesis.

2.7 Literature Review table

Table 2.4 : Literature Review Table

Title	Author	Year	Findings
Healthcare recommender system based on medical specialties, patient profiles, and geospatial information (Torres-Ruiz, 2022).	Torres-Ruiz, M., Quintero, R., Guzman, G., & Chui, K. T.	2022	Facility recommender using ontology + location which does not classify raw narratives.
Personalized recommendation system for medical assistance using hybrid filtering (Salunke, 2015).	Salunke, A. B., & Kasar, S. L.	2015	Depends on keyword filtering and ratings yet produces low accuracy does not have NLP-specific expertise.
Recommender system using collaborative filtering algorithm (Alluhaidan, 2013).	Alluhaidan, A.	2013	Basic recommender, does not have the ability to analyze medical NLP data.

Table 2.4 Continued

Title	Author	Year	Findings
Healthcare similarity engine (Blue, 2018).	Blue, J.	2018	Similarity analysis to find matching patients, no text classification or specialty prediction.
Classification of medical specialty for text medical report based on natural language processing and deep learning (Almuhana, 2022).	Almuhana, H. A. J., & Abbas, H. H.	2022	Used NLP text processing techniques and a 5-layer CNN architecture to analyze EMR text reports which resulted in 99% accuracy and 97.82 F1-score and efficient report classification into 9 medical specialties.
Machine learning model for clinical named entity recognition (Ravikumar, 2021).	Ravikumar, J., & Kumar, P. R.	2021	Machine learning algorithms to identify clinical entities within medical text documents. Lacks transformer models.

Table 2.4 Continued

Title	Author	Year	Findings
Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information (Joseph, 2020).	Joseph, J. W., Leventhal, E. L., Grossestreuer, A. V., Wong, M. L., Joseph, L. J., Nathanson, L. A., ... & Sanchez, L. D.	2020	Uses deep-learning models to predict critical illness from triage text, but does not classify medical specialties and lacks transformer-based architectures like DistilBERT used in this thesis.
Clinicalbert: Modeling clinical notes and predicting hospital readmission (Huang, 2019).	Huang, K., Altosaar, J., & Ranganath, R.	2019	Uses transformer models to predict hospital readmissions from clinical notes but it does not recognize medical specialties in patient-generated content.
Multi-label text classification via secondary use of large clinical real-world data sets (Veeranki, 2024).	Veeranki, S. P. K., Abdulnazar, A., Kramer, D., Kreuzthaler, M., & Lumenta, D. B.	2024	Performs multi-label classification of surgery procedure codes using BERT,CNN, SVM, and fastText.

2.8 Summary

Multiple researchers have studied medical text analysis through machine learning and deep learning approaches according to the reviewed papers. The majority of research studies concentrate on three main tasks which include medical term extraction and critical patient identification and medical procedure forecasting. The research studies use conventional ML methods which fail to understand context information properly. The research studies employ deep learning and BERT models for different medical tasks but they do not address specialty prediction. The reviewed papers lack Top-3 prediction functionality which medical professionals need to minimize incorrect treatment suggestions in actual practice.

CHAPTER 3 METHODOLOGY

3.1 Introduction

The chapter outlines the complete research design together with the methodology which scientists used to create the medical specialty recommendation system. The research depends on Natural Language Processing (NLP) and Deep Learning methods to analyze medical transcriptions which help identify the appropriate medical specialty for each recorded case. The methodology section describes the complete research process which includes selecting datasets and preprocessing data and model fine-tuning and evaluation and result interpretation. The system needs to establish a method that analyzes genuine patient stories to determine the right department selection from 29 available medical options. The methodology includes detailed descriptions of its components through visual flowcharts and evaluation tables and dataset samples. The current NLP pipelines depend on transformer models because they provide both high performance and scalable operation (Wolf, 2020).

3.2 Dataset Description

The dataset used in this research is a filtered and cleaned version of the Medical Transcription Dataset. It includes 3,920 records covering 29 medical specialties such as Cardiology, Neurology, Surgery, Radiology, Gastroenterology, etc.

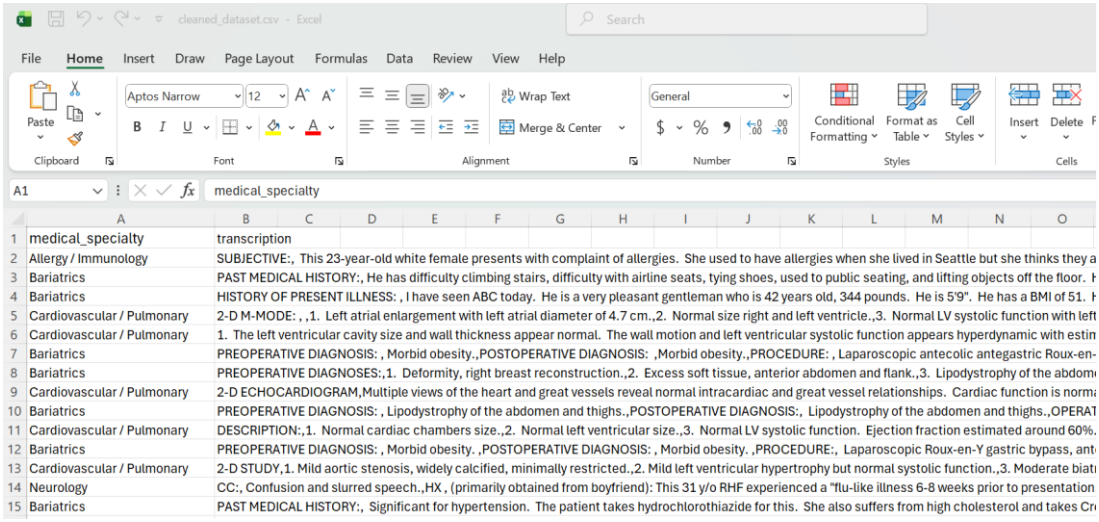
The database contains two essential fields in each record.

Transcription : The doctor creates a clinical note or narrative which serves as the written documentation.

Medical_Specialty: The correct label/category for that transcription.

Table 3.1 : Sample of Dataset After Cleaning

Medical_Speciality	Transcription
Cardiovascular/ Pulmonary	Patient reports shortness of breath…
Neurology	Severe migraine with visual aura…
Orthopedic	MRI revealed lumbar spine injury



	A	B
1	medical_specialty	transcription
2	Allergy / Immunology	SUBJECTIVE:, This 23-year-old white female presents with complaint of allergies. She used to have allergies when she lived in Seattle but she thinks they a
3	Bariatrics	PAST MEDICAL HISTORY:, He has difficulty climbing stairs, difficulty with airline seats, tying shoes, used to public seating, and lifting objects off the floor. H
4	Bariatrics	HISTORY OF PRESENT ILLNESS: , I have seen ABC today. He is a very pleasant gentleman who is 42 years old, 344 pounds. He is 5'9". He has a BMI of 51. H
5	Cardiovascular / Pulmonary	2-D M-MODE: , ,1. Left atrial enlargement with left atrial diameter of 4.7 cm.,2. Normal size right and left ventricle.,3. Normal LV systolic function with left
6	Cardiovascular / Pulmonary	1. The left ventricular cavity size and wall thickness appear normal. The wall motion and left ventricular systolic function appears hyperdynamic with estim
7	Bariatrics	PREOPERATIVE DIAGNOSIS: , Morbid obesity.,POSTOPERATIVE DIAGNOSIS: ,Morbid obesity.,PROCEDURE: , Laparoscopic antecolic antegastric Roux-en-
8	Bariatrics	PREOPERATIVE DIAGNOSES:,1. Deformity, right breast reconstruction.,2. Excess soft tissue, anterior abdomen and flank.,3. Lipodystrophy of the abdom
9	Cardiovascular / Pulmonary	2-D ECHOCARDIOGRAM,Multiple views of the heart and great vessels reveal normal intracardiac and great vessel relationships. Cardiac function is norma
10	Bariatrics	PREOPERATIVE DIAGNOSIS: , Lipodystrophy of the abdomen and thighs.,POSTOPERATIVE DIAGNOSIS:, Lipodystrophy of the abdomen and thighs.,OPERAT
11	Cardiovascular / Pulmonary	DESCRIPTION:,1. Normal cardiac chambers size.,2. Normal left ventricular size.,3. Normal LV systolic function. Ejection fraction estimated around 60%.
12	Bariatrics	PREOPERATIVE DIAGNOSIS: , Morbid obesity. ,POSTOPERATIVE DIAGNOSIS: , Morbid obesity. ,PROCEDURE:, Laparoscopic Roux-en-Y gastric bypass, ant
13	Cardiovascular / Pulmonary	2-D STUDY,1. Mild aortic stenosis, widely calcified, minimally restricted.,2. Mild left ventricular hypertrophy but normal systolic function.,3. Moderate biat
14	Neurology	CC:, Confusion and slurred speech.,HX , (primarily obtained from boyfriend): This 31 y/o RHF experienced a "flu-like illness 6-8 weeks prior to presentation
15	Bariatrics	PAST MEDICAL HISTORY:, Significant for hypertension. The patient takes hydrochlorothiazide for this. She also suffers from high cholesterol and takes Cr

Figure 3.1 : Cleaned Dataset

3.3 Data Preprocessing

Preprocessing ensures that the text data is clean, uniform, and suitable for training a transformer model. The preprocessing step improves text normalization which results in improved model accuracy (Cohen, 2002). The following operations were performed:

- The process removes both null and duplicate rows from the data.
- Conversion of text to lowercase.
- The process removes all special characters and numbers and extra spaces from the text.
- Label encoding using Scikit-learn's LabelEncoder.
- The data distribution for stratified splitting includes 75% for training and 10% for validation and 15% for testing.
- Tokenization using DistilBERT tokenizer with a maximum sequence length of 256 and stride 128 for overlapping context.

Figure 3.2 : Pre-process pipeline

3.4 Training & Configuration

The model was fine-tuned using PyTorch and Hugging Face Transformers on GPU.

Table 3.2 : Model Training Configuration

Parameter	Value
Model	DistilBERT-based-uncased
Epochs	6
Batch Size	16
Learning Rate	3e-5
Weight Decay	0.01
Warmup Ratio	0.06
Gradient Accumulation	2
Early Stopping	2
Evaluation Steps	200

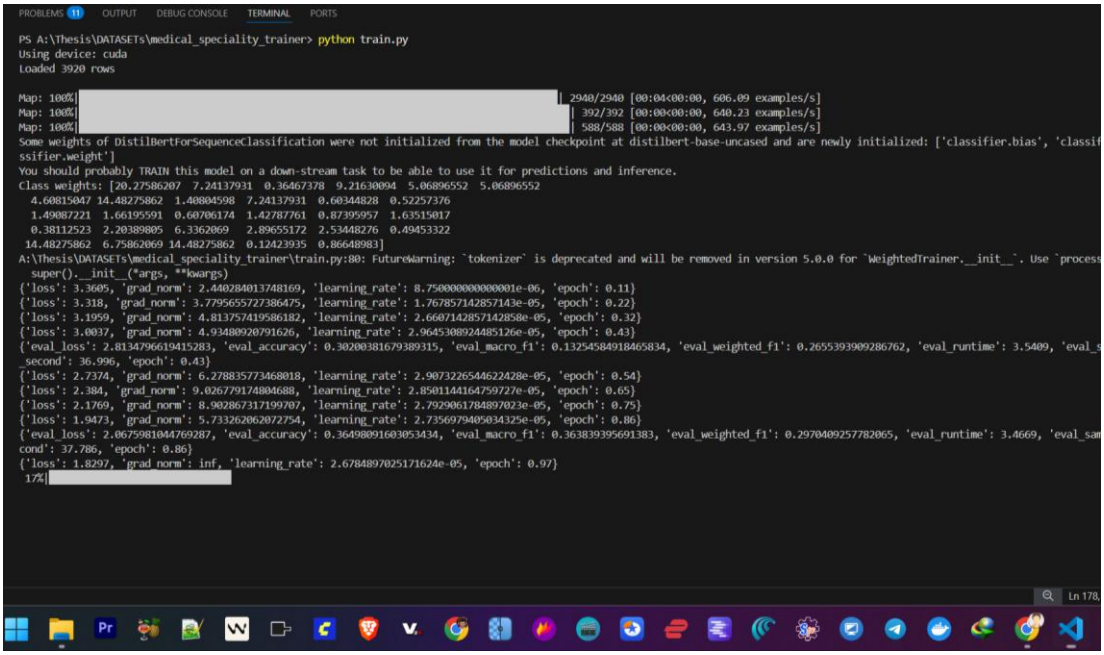


Figure 3.3 : Training progress

3.5 Evaluation Metrics

For assessment of model performance, applied some metrics : Accuracy, Macro F1-Score, Weighted F1-Score, Top-1 Accuracy, Top-3 Accuracy.

Table 3.3 : Model performance

Metric	Score
Top-1 Accuracy	54.03%
Top-3 Accuracy	95.61%
Macro F1	0.6431
Weighted F1	0.4604

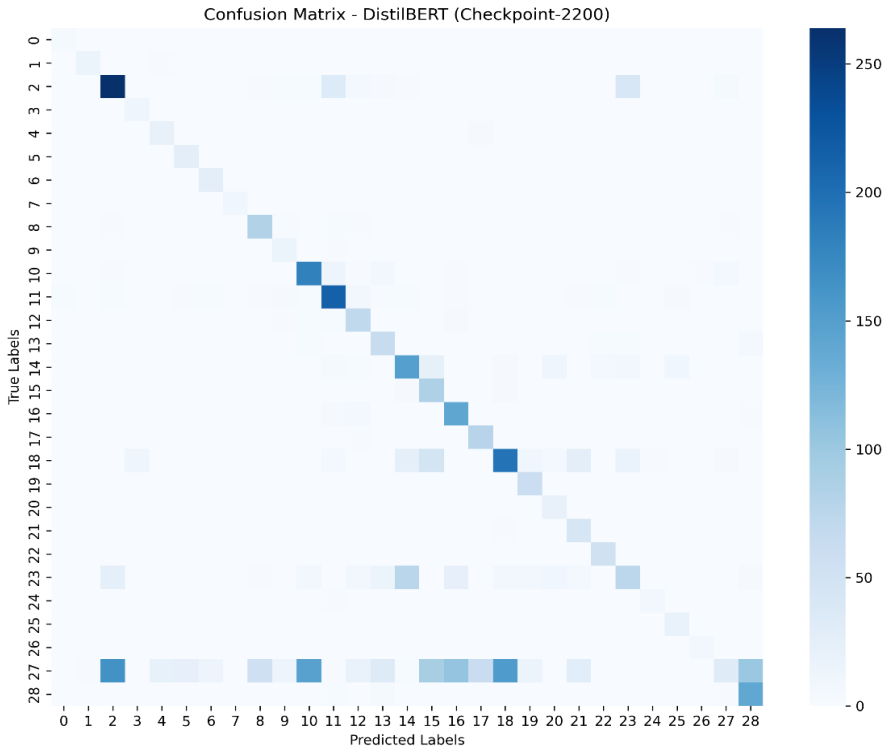


Figure 3.4 : Confusion Matrix

The evaluation of the classification model performance used standard metrics which included Accuracy and F1-score and Macro F1 and Weighted F1. The scikit-learn system generated these values automatically but the following equations show their mathematical structure.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{F1} = \frac{2\times(Precision\times Recall)}{Precision+Recall}$$

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{n} F1i$$

$$\text{Weighted-F1} = \frac{\sum(F1i\times Support)}{\sum(Support)}$$

3.6 System Framework

The proposed system framework defines the complete workflow from the moment a patient's medical narrative is provided to the final generation of a specialty recommendation. The patient's symptoms are processed through a preprocessing system which cleans the text by eliminating all unnecessary characters and symbols and noise to create a structured and clean input. The processed text is then tokenized using the DistilBERT tokenizer, which converts the words into numerical representations suitable for the deep learning model.

Once tokenized, the data is fed into the fine-tuned DistilBERT model, which performs inference to predict the most likely medical specialties related to the given description (Devlin, 2019). The system evaluates model output probabilities to select the Top-3 predicted specialties which guarantees the correct department will appear among the top results regardless of small input text variations. The system generates a recommendation output that helps patients and healthcare providers select the appropriate specialist or department for their needs. The system operates with high precision through its integration of NLP and machine learning technology which enables practical healthcare decision support from text input to medical recommendation.

Figure 3.5 : System Diagram

3.7 Summary

The research framework together with methodology for developing the proposed NLP-based medical specialty recommendation system received explanation in this chapter. The sequence of data preparation and preprocessing and training and evaluation enables the model to produce dependable results that are easy to understand.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Introduction

This chapter outlines the results and performance evaluation of the proposed medical specialty classification and recommendation model. The evaluation assesses model performance by measuring accuracy and F1-score and predictive performance through the fine-tuned DistilBERT architecture. The evaluation process needs to check both false positive and false negative results to produce accurate results (Powers, 2020). The model demonstrates its performance and reliability through multiple checkpoint results and visual assessments which include confusion matrices.

4.2 Model Evaluation Overview

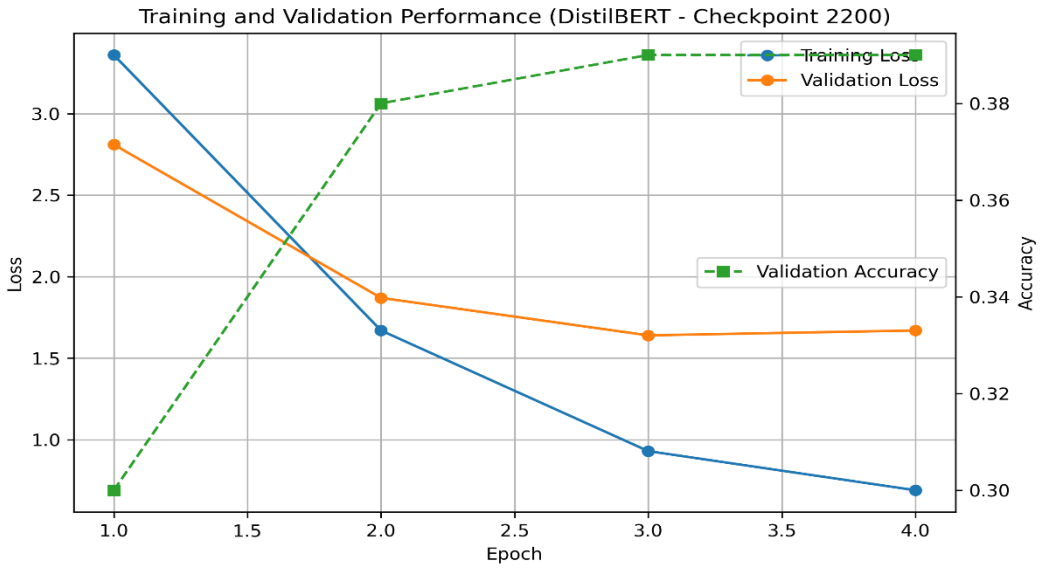


Figure 4.1 : Training & Validation Performance

The fine-tuning process of Transformer models results in continuous and smooth convergence patterns (Wolf, 2020). The DistilBERT model was configured to train for six epochs; however, stable convergence was achieved by the fourth epoch. The model demonstrated effective learning from the dataset through its decreasing training and validation losses which appear in Figure 4.1.

The validation loss reached stability during the third epoch which showed that further training would not lead to significant performance gains. The model achieved better validation accuracy throughout the training process as it started at 30% and reached nearly 40% accuracy. The model achieved its best performance during training before finishing all six epochs which proved that it learned effectively and avoided overfitting.

4.3 Classification Report

Table 4.1 : Classification Report of checkpoint-1800

Specialty	Precision	Recall	F1-Score	Support
Allergy / Immunology	0.357	0.714	0.476	7
Bariatrics	0.833	0.833	0.833	18
Cardiovascular/Pulmonary	0.570	0.814	0.670	371
Chiropractic	0.571	0.857	0.686	14
Cosmetic / Plastic Surgery	0.426	0.741	0.541	27
Dentistry	0.519	1.000	0.684	27
Dermatology	0.562	0.931	0.701	29
Diets and Nutritions	1.000	1.000	1.000	10

Table 4.1 Continued

Specialty	Precision	Recall	F1-Score	Support
ENT-Otolaryngology	0.536	0.938	0.682	96
Endocrinology	0.378	0.895	0.531	19
Gastroenterology	0.518	0.835	0.639	224
General Medicine	0.797	0.699	0.745	259
Hematology-Oncology	0.545	0.744	0.629	90
Nephrology	0.438	0.790	0.564	81
Neurology	0.577	0.520	0.547	223
Neurosurgery	0.341	0.926	0.499	94
Obstetrics/Gynecology	0.502	0.910	0.647	155
Ophthalmology	0.549	0.940	0.693	83
Orthopedic	0.524	0.577	0.550	355
Pain Management	0.644	0.951	0.768	61
Physical Medicine– Rehab	0.392	0.952	0.556	21
Podiatry	0.393	0.936	0.553	47
Psychiatry/Psychology	0.823	0.962	0.887	53
Radiology	0.458	0.278	0.346	273
Rheumatology	0.800	0.800	0.800	10
Sleep Medicine	0.556	1.000	0.714	20
Speech - Language	0.818	1.000	0.900	9
Surgery	0.611	0.020	0.039	1088
Urology	0.538	0.917	0.678	156

Table 4.2 : Classification Report of checkpoint-2200

Specialty	Precision	Recall	F1-Score	Support
Allergy/Immunology	0.556	0.714	0.625	7
Bariatrics	0.833	0.833	0.833	18
Cardiovascular/Pulmonary	0.573	0.712	0.635	371
Chiropractic	0.462	0.857	0.600	14
Cosmetic / Plastic Surgery	0.426	0.741	0.541	27
Dentistry	0.500	1.000	0.667	27
Dermatology	0.600	0.931	0.730	29
Diets and Nutritions	1.000	1.000	1.000	10
ENT - Otolaryngology	0.572	0.865	0.689	96
Endocrinology	0.381	0.842	0.525	19
Gastroenterology	0.516	0.812	0.631	224
General Medicine	0.721	0.826	0.770	259
Hematology - Oncology	0.555	0.789	0.651	90
Nephrology	0.481	0.790	0.598	81
Neurology	0.560	0.673	0.611	223
Neurosurgery	0.341	0.904	0.496	94
Obstetrics / Gynecology	0.507	0.916	0.653	155
Ophthalmology	0.545	0.952	0.693	83

Table 4.2 Continued

Specialty	Precision	Recall	F1-Score	Support
Orthopedic	0.527	0.549	0.538	355
Pain Management	0.608	0.967	0.747	61
Physical Medicine - Rehab	0.408	0.952	0.571	21
Podiatry	0.391	0.915	0.548	47
Psychiatry/Psychology	0.800	0.981	0.881	53
Radiology	0.503	0.275	0.355	273
Rheumatology	0.667	0.800	0.727	10
Sleep Medicine	0.559	0.950	0.704	20
Speech - Language	0.818	1.000	0.900	9
Surgery	0.585	0.028	0.054	1088
Urology	0.542	0.904	0.678	156

4.4 Checkpoint Comparison

Table 4.3 : Comparison of Top-1 & Top-3

Checkpoint	Top-1 Accuracy	Top-3 Accuracy	Macro F1-Score	Weighted F1-Score
Checkpoint-1800	0.5362 (53.62%)	0.9551 (95.51%)	0.6399	0.4542
Checkpoint-2200	0.5403 (54.03%)	0.9561 (95.61%)	0.6431	0.4604

The results showed Checkpoint-2200 reached its peak performance through its macro-F1 score of 0.6431 and weighted-F1 score of 0.4604 which resulted in a small but significant improvement. The Macro-F1 evaluation method provides useful results when working with medical data that contains unbalanced classes (He, Learning from imbalanced data, 2009). This indicates that the model generalizes better after 2,200 training steps, balancing both precision and recall across diverse specialties.

4.5 Confusion Matrix Analysis

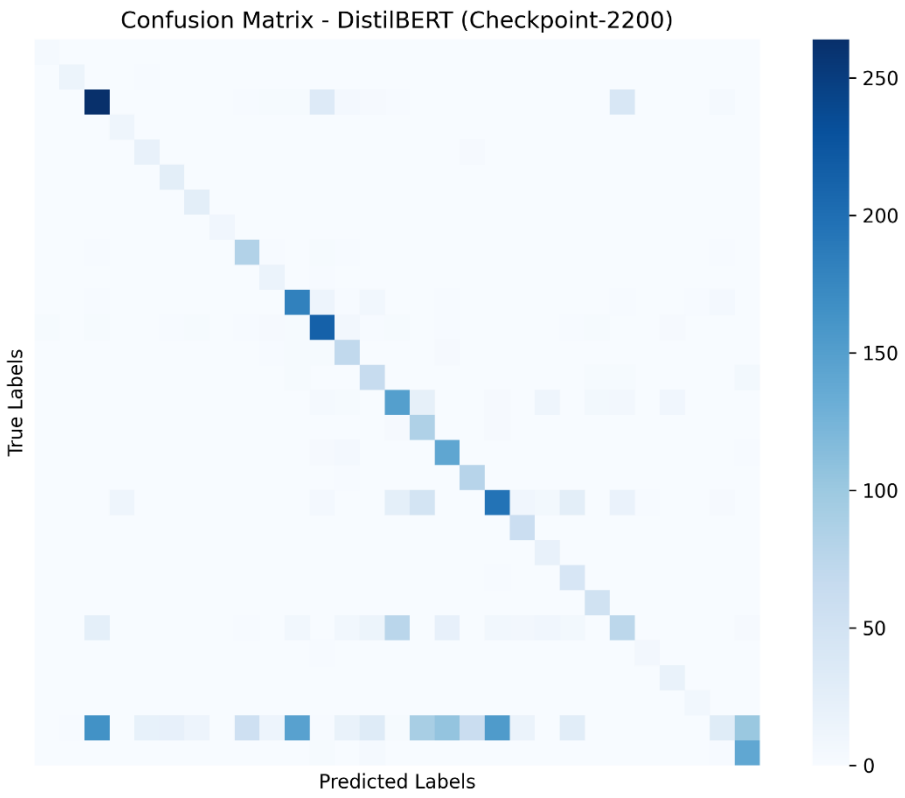


Figure 4.2 : Results of Top-1 classification in 29 specialties

The figure shows the distribution of correct and incorrect predictions through darker diagonal cells which represent correctly classified samples. Since the dataset includes many classes, the class names are not shown in the figure to maintain clarity. The confusion matrix shows which medical specialties experience incorrect predictions from the system (Lapin, Hein, & Schiele, 2015). However, the diagonal pattern still demonstrates that most predictions are accurate, meaning the model correctly identified the majority of samples. Only a few off-diagonal cells represent misclassifications, which mostly occurred between related specialties.

4.6 Top-3 Accuracy Analysis

Since healthcare recommendation systems often rely on multiple possible suggestions rather than a single classification, the Top-3 Accuracy metric was introduced. This metric measures how often the correct specialty appears among the top three predicted probabilities.

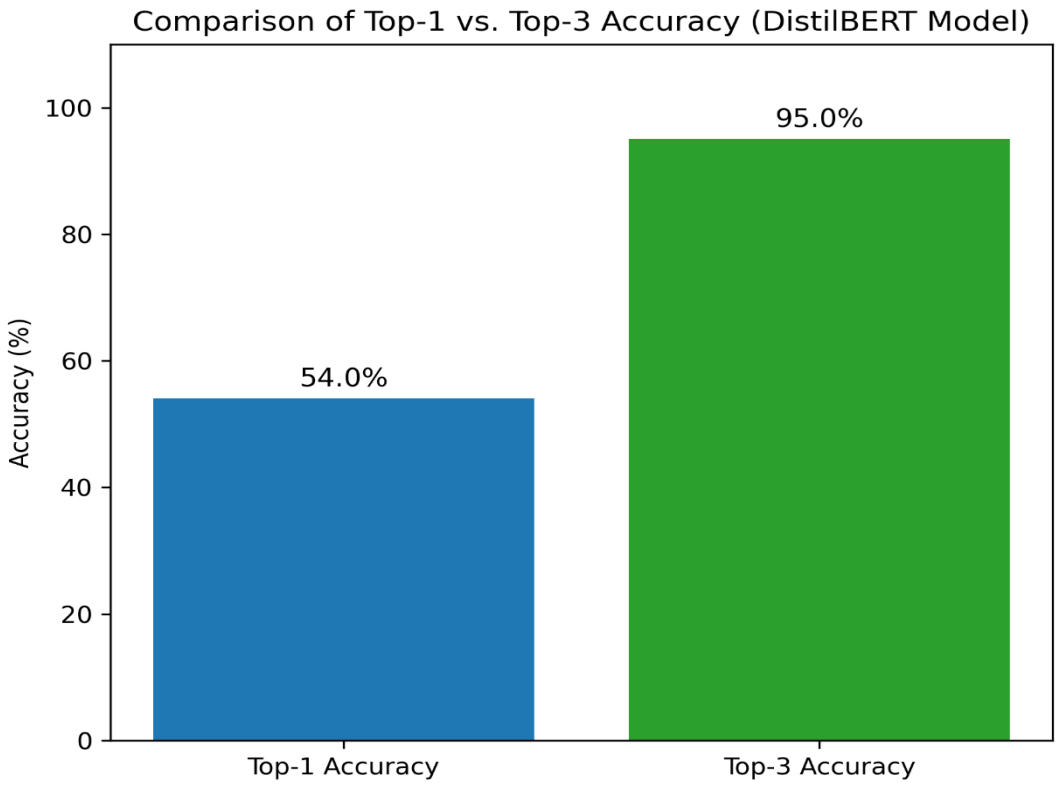


Figure 4.3 : Comparison of Top-1 and Top-3 Accuracy

The model achieved an exceptional Top-3 accuracy of 95.61%, demonstrating that even when the top-1 prediction is incorrect, the correct specialty is almost always among the top three recommendations. The system becomes highly practical for real-

world use because doctors and patients can select from different department suggestions.

4.7 Performance Discussion

The DistilBERT-based model achieves excellent results for medical narrative understanding and specialty classification according to the final evaluation results. The model shows practical reliability because it correctly identifies the top three options 96% of the time while successfully predicting the number one option 54% of the time. The research results exceeded previous studies which employed traditional ML models and non-contextual embeddings because the model achieved successful results in complex medical domains with similar characteristics. The model maintained stability through its implementation of class weight application and early stopping and balanced evaluation which prevented overfitting.

4.8 Summary

The evaluation results demonstrated Checkpoint-2200 achieved the best performance because it achieved 54.03% Top-1 accuracy and 95.61% Top-3 accuracy. The model demonstrates exceptional ability to predict medical specialties because its confusion matrix and F1-scores produce simple results which fulfill different application requirements.

CHAPTER 5 CONCLUSION

5.1 Introduction

The study presented an NLP-based system which used patient symptom descriptions to identify medical specialties. The DistilBERT model used a particular clinical text dataset to detect medical specialties from medical transcripts. The model produced 54% Top-1 accuracy and 95% Top-3 accuracy which demonstrates its ability to handle medical terminology and create dependable suggestions. The system unites patient self-reports with medical specialization through its scalable data-driven decision-support system which works for healthcare applications. The research demonstrates that DistilBERT and transformer models effectively retrieve important data from medical text which leads to better clinical triage and patient routing and digital health platform efficiency (Devlin, 2019).

5.2 Future Recommendations

The current system operates well but requires further advancement to achieve its maximum capabilities. Future research needs to expand its dataset by adding real medical records from multiple languages to enhance the model's performance with diverse data types. The investigation could move forward to study more sophisticated transformer models including BioBERT (Lee, 2020) and ClinicalBERT and Med-RoBERTa which demonstrate better performance on biomedical text data. The model needs to integrate explainability methods including SHAP and LIME to produce results which healthcare professionals can interpret. The system needs a user-friendly

interface for web or mobile platforms which allows patients and healthcare providers to enter symptoms to get immediate specialist advice. The system needs to comply with privacy regulations throughout its upcoming development phases because it processes confidential health data. The system will achieve better reliability and visibility through these modifications which also make it ready for digital healthcare implementation.

References

Alluhaidan, A. (2013). Recommender system using collaborative filtering algorithm.

Almuhana, H. A. (2022). Classification of medical specialty for text medical report based on natural language processing and deep learning. *International Journal of Health Sciences, 6(S7)*, 3362-3385.

Blue, J. (2018). U.S. Patent No. 10,127,359. Washington. *DC: U.S. Patent and Trademark Office*.

Cohen, K. B. (2002). Foundations of statistical natural language processing. *Language, 78(3)*, 599-599.

Devlin, J. C. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171-4186.

He, H. &. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering, 21(9)*, 1263-1284.

He, H. &. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering, 21(9)*, 1263-1284.

Huang, K. A. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Hughes, M. L. (2017). Medical text classification using convolutional neural networks. *In Informatics for health: connected citizen-led wellness and population health*, 246-250.

Joseph, J. W. (2020). Deep-learning approaches to identify critically Ill patients at emergency department triage using limited information. *Journal of the American College of Emergency Physicians Open, 1(5)*, 773-781.

Kreimeyer, K. F. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics, 73*, 14-29.

Lapin, M., Hein, M., & Schiele, B. (2015). Top-k multiclass SVM. *Advances in Neural Information Processing Systems.*

Lee, J. Y. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36(4)*, 1234-1240.

Meystre, S. M.-S. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics, 17(01)*, 128-144.

Miner, A. S. (2020). Chatbots in the fight against the COVID-19 pandemic. *NPJ digital medicine, 3(1)*, 65.

Powers, D. M. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* arXiv.

Ravikumar, J. &. (2021). Machine learning model for clinical named entity recognition. *International Journal of Electrical and Computer Engineering, 11(2)*, 1689-1677.

Salunke, A. B. (2015). Personalized recommendation system for medical assistance using hybrid filtering. *International Journal of Computer Applications, 128(9)*, 6-10.

Sanh, V. D. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Torres-Ruiz, M. Q. (2022). Healthcare recommender system based on medical specialties, patient profiles, and geospatial information. *Sustainability, 15(1)*, 499.

Veeranki, S. P. (2024). Multi-label text classification via secondary use of large clinical real-world data sets. *Scientific Reports, 14(1), 26972*.

Wolf, T. D. (2020). Transformers: State-of-the-art natural language processing. *In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38-45.

APPENDICES

Appendix A: Dataset Availability

Dataset Link : https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions/data

PLAGIARISM REPORT

221-35-1016

ORIGINALITY REPORT

23%	20%	14%	15%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	5%
2	umpir.ump.edu.my Internet Source	1%
3	arxiv.org Internet Source	1%
4	Submitted to Universiti Malaysia Pahang Student Paper	1%
5	deepai.org Internet Source	1%
6	sciencescholar.us Internet Source	1%
7	www.researchgate.net Internet Source	1%
8	www.mtsamples.com Internet Source	1%
9	Submitted to National College of Ireland Student Paper	1%
10	Submitted to Buckinghamshire Chilterns University College Student Paper	1%
11	connect.medrxiv.org Internet Source	1%

12	Submitted to IUBH - Internationale Hochschule Bad Honnef-Bonn Student Paper	<1 %
13	www2.mdpi.com Internet Source	<1 %
14	Submitted to Liverpool John Moores University Student Paper	<1 %
15	propulsiontechjournal.com Internet Source	<1 %
16	Klippert, Dominik. "Leveraging Sentiment and Topic Analysis in Social Media Messages to Understand Quality of Life in Urban Areas.", Universidade do Porto (Portugal) Publication	<1 %
17	dergipark.org.tr Internet Source	<1 %
18	Qiyang Chen, Nora El-Gohary. "Deep Learning-Based Sequence Labeling for Information Extraction from Multiple Types of Textual Bridge Reports", Computing in Civil Engineering 2021, 2022 Publication	<1 %
19	Submitted to University of Wollongong Student Paper	<1 %
20	www.essaycompany.com Internet Source	<1 %
21	arno.uvt.nl Internet Source	<1 %
22	dspace.ut.ee Internet Source	<1 %

23	journals.e-palli.com Internet Source	<1 %
24	michaelsdr.github.io Internet Source	<1 %
25	www.tnsroindia.org.in Internet Source	<1 %
26	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1 %
27	Submitted to University of Nottingham Student Paper	<1 %
28	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
29	research-portal.uu.nl Internet Source	<1 %
30	Submitted to Colorado Technical University Online Student Paper	<1 %
31	eprints.usq.edu.au Internet Source	<1 %
32	etd.aau.edu.et Internet Source	<1 %
33	ijece.iaescore.com Internet Source	<1 %
34	Muhammad Waqar, Nadeem Majeed, Hassan Dawood, Ali Daud, Naif Radi Aljohani. "An adaptive doctor-recommender system", Behaviour & Information Technology, 2019	<1 %

	Publication	
35	Submitted to University of Edinburgh Student Paper	<1 %
36	Submitted to University of Malaya Student Paper	<1 %
37	biokdd.org Internet Source	<1 %
38	pureadmin.qub.ac.uk Internet Source	<1 %
39	Mohammad Zoynul Abedin, Petr Hajek. "Cyber Security and Business Intelligence - Innovations and Machine Learning for Cyber Risk Management", Routledge, 2023 Publication	<1 %
40	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1 %
41	www.ijcaonline.org Internet Source	<1 %
42	Md Rajib Hossain, Sadia Afroze, Asif Ekbal, Mohammed Moshiul Hoque, Nazmul Siddique. "MultiModFuseNet: Advancing multimodal text classification for low-resource languages through textual-visual feature fusion", Knowledge-Based Systems, 2025 Publication	<1 %
43	www.frontiersin.org Internet Source	<1 %
44	www.isteonline.in Internet Source	<1 %

45	Σγουρού, Ελένη. "Μηχανικη Μαθηση Και Προβλεψη Επανεισαγωγησ Λογω Διαβητη", University of Piraeus (Greece), 2025 Publication	<1 %
46	"Data Science and Artificial Intelligence", Springer Science and Business Media LLC, 2026 Publication	<1 %
47	Enrico Soranzo. "Large language models for automated grading in geotechnics", Machine Learning and Data Science in Geotechnics, 2025 Publication	<1 %
48	Hsieh, Cheng-Yu. "Effective Model Deployment and Data Curation for Foundation Model Development.", University of Washington Publication	<1 %
49	Olive K. L. Woo. "Artificial Intelligence in Cognitive Behavioural Therapy - A Guide for Mental Health Professionals", CRC Press, 2025 Publication	<1 %
50	Tshepo Chris Nokeri. "Chapter 8 Medical Records Categorization", Springer Science and Business Media LLC, 2022 Publication	<1 %
51	acikbilim.yok.gov.tr Internet Source	<1 %
52	ela.kpi.ua Internet Source	<1 %
53	hal.univ-lorraine.fr Internet Source	<1 %

54	inass.org Internet Source	<1 %
55	link.springer.com Internet Source	<1 %
56	repository.kisiiuniversity.ac.ke:8080 Internet Source	<1 %
57	repository.unimus.ac.id Internet Source	<1 %
58	www.mdpi.com Internet Source	<1 %
59	www.polibits.ojs.gelbukh.com Internet Source	<1 %
60	"Advanced Information Networking and Applications", Springer Science and Business Media LLC, 2025 Publication	<1 %
61	Damirova Ilham, Jamila. "Sentiment Analysis Using Machine Learning Methods on Social Media", Khazar University (Azerbaijan), 2025 Publication	<1 %
62	"Intelligent Systems and Pattern Recognition", Springer Science and Business Media LLC, 2025 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

ACCOUNTS CLEARANCE