Clinical Notes Classification Using Deep Learning

ORGANIZATION: DENVEY EDUGROW

PROGRAM: AI/ Deep learning / AI in Healthcare

Submission Date
10/01/2026

Submitted by:
NAME: Salman Firdous
Roll No : 242025062

DECLARATION

I hereby declare that this project titled **“Clinical Notes Classification Using Deep Learning”** is my original work carried out under the guidance of the academic team of **Denvey EduGrow**. This project has not been submitted elsewhere for any academic degree or certification

Date: 09/01/2026
Place: Guwahati

Signature
Salman Firdous
Roll no:242025062

Acknowledgement

I would like to express my sincere gratitude to **Denvey EduGrow** for providing the opportunity and guidance to complete this project. I also thank the mentors and instructors for their continuous support and valuable feedback throughout the project duration.

Thank You

Salman Firdous

Roll no: 242025062

Abstract

Clinical notes contain unstructured medical text such as diagnosis details, prescriptions, lab reports, and discharge summaries. Manual classification of these notes is time-consuming and error-prone. This project focuses on developing a **deep learning–based clinical notes classification system** using Natural Language Processing techniques. A Bidirectional LSTM model is trained to automatically classify clinical notes into predefined categories, improving efficiency and accuracy in healthcare data management.

TABLE OF CONTENTS

1. Introduction

Healthcare systems generate a large amount of clinical data in the form of diagnosis notes, prescriptions, lab reports, discharge summaries, and progress notes. These clinical notes contain important medical information but are usually written in free-text format, making them difficult to organize and analyze manually. Managing such unstructured data requires significant time and effort from healthcare professionals.

Manual classification of clinical notes is not only time-consuming but also prone to human errors and inconsistencies. With the increasing volume of patient data, there is a growing need for automated systems that can efficiently process and categorize clinical text. Artificial Intelligence and Natural Language Processing (NLP) provide effective solutions for handling unstructured medical data.

This project focuses on developing a deep learning–based system for automatic clinical notes classification. By using NLP techniques and a Bidirectional LSTM model, the system learns patterns from medical text and classifies clinical notes into predefined categories such as diagnosis, prescription, lab reports, progress notes, and discharge summaries. The proposed system helps improve healthcare data organization and supports efficient clinical information management.

2.Problem Statement

Clinical notes such as diagnosis records, prescriptions, lab reports, discharge summaries, and progress notes are mostly written in unstructured text form. Manually organizing and classifying these clinical notes is a time-consuming process and requires significant human effort. It is also prone to errors and inconsistencies, especially when handling large volumes of healthcare data.

With the rapid growth of electronic health records, there is a need for an automated and reliable system that can accurately classify clinical notes into appropriate categories. An effective solution should reduce manual workload, improve accuracy, and support efficient healthcare data management. This project addresses this need by applying deep learning and Natural Language Processing techniques to automate clinical notes classification.

3.Objectives

The objective of this project is to design and implement a deep learning–based system for automatic classification of clinical notes into predefined medical categories. The project aims to apply Natural Language Processing techniques to preprocess clinical text and train a Bidirectional LSTM model to learn meaningful patterns from the data. Another objective is to evaluate the performance of the model using standard metrics such as accuracy, precision, recall, and F1-score, with the goal of improving the efficiency and accuracy of managing unstructured healthcare text data.

4. Dataset Description

The dataset used in this project consists of textual clinical notes and their corresponding category labels. These clinical notes include information such as diagnosis details, prescriptions, laboratory reports, progress notes, and discharge summaries. Each record in the dataset contains a clinical note in text form along with a label representing its category.

The dataset is stored in CSV format and was synthetically created for academic and learning purposes. Before training the deep learning model, the dataset was preprocessed to remove unnecessary characters and noise, ensuring better model performance. Since the dataset size is small, it is included directly in the GitHub repository.

5. Methodology

The methodology of this project follows a systematic approach to classify clinical notes using deep learning techniques. Initially, the clinical notes dataset is collected and loaded for processing. Text preprocessing is performed to clean the data by converting text to lowercase and removing special characters and extra spaces. The cleaned text is then converted into numerical form using tokenization, and padding is applied to ensure uniform input length.

Next, the target labels are encoded, and the dataset is divided into training and testing sets. A Bidirectional LSTM deep learning model is designed and trained on the processed data to learn contextual patterns in clinical text. Finally, the trained model is evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score to measure its effectiveness in classifying clinical notes.

6. Tools & Technologies Used

The following tools and technologies were used for the successful implementation of this project. The programming was carried out using **Python**, which provides strong support for machine learning and deep learning applications. Data processing and analysis were performed using **Pandas** and **NumPy**. The deep learning model was developed using **TensorFlow** and **Keras**, while **Scikit-learn** was used for preprocessing tasks and performance evaluation.

The project was implemented using **Google Colab** as the development platform. **GitHub** was used for version control and project repository management

7. Deep Learning Model Used

In this project, a **Bidirectional Long Short-Term Memory (Bi-LSTM)** deep learning model is used for clinical notes classification. LSTM networks are well-suited for processing sequential text data as they can learn long-term dependencies in sentences. The Bidirectional LSTM processes the text in both forward and backward directions, allowing the model to capture better contextual information from clinical notes. This model improves classification accuracy by understanding the meaning and structure of medical text more effectively.

8. Performance Evaluation

The performance of the proposed clinical notes classification system is evaluated using standard evaluation metrics. These metrics include **accuracy**, **precision**, **recall**, and **F1-score**, which are commonly used to measure the effectiveness of classification models. Accuracy measures the overall correctness of predictions, while precision and recall evaluate the model's ability to correctly classify each category. The F1-score provides a balanced measure of precision and recall. These metrics help in analyzing the reliability and efficiency of the deep learning model.

9. Results and Discussion

The experimental results show that the deep learning–based clinical notes classification system is able to classify clinical text into predefined categories with reasonable accuracy. The Bidirectional LSTM model successfully learns contextual patterns from clinical notes and performs well in distinguishing between different types of medical text such as diagnosis, prescription, lab reports, progress notes, and discharge summaries.

The results indicate that the use of Natural Language Processing techniques combined with deep learning improves the efficiency of handling unstructured clinical data. Although the performance depends on the size and quality of the dataset, the model demonstrates the potential of deep learning approaches in automating clinical text classification and supporting effective healthcare data management.

10. Conclusion

This project successfully demonstrates the application of deep learning and Natural Language Processing techniques for the automatic classification of clinical notes. The developed Bidirectional LSTM model effectively categorizes clinical text into predefined medical classes, reducing manual effort and improving data organization. The system shows that deep learning can play an important role in managing unstructured healthcare data and supporting efficient clinical information management.

11. Future Scope

The proposed clinical notes classification system can be further enhanced in several ways. The model can be trained on larger and real-world clinical datasets to improve accuracy and generalization. Advanced deep learning models such as BERT and transformer-based architectures can be implemented for better understanding of medical text. The system can also be integrated with Electronic Health Record (EHR) systems and deployed as a web-based application to support real-time clinical data management.

12. References

1. TensorFlow Documentation – Deep Learning Framework

2. Keras Documentation – Neural Network API

3. Scikit-learn Documentation – Machine Learning Tools

4. Research papers on Natural Language Processing in Healthcare

5. AI and Machine Learning applications in Clinical Text Analysis