Transforming Medical Transcripts: NLP for Enhanced Healthcare Data Management

Anna Dominic
amd9200@nyu.edu

Siri Desiraju
scd4156@nyu.edu

Sri Raksha Gubbihal
sg7374@nyu.edu

Abstract

This project develops a methodology to convert unstructured EHR data into structured formats, employing diverse classification models for accurate categorization. The integration of the T5 model is pivotal, transforming unstructured transcriptions into structured data, thereby enhancing healthcare data management. Our approach, validated through experiments, shows improved classification and robust data transformation, aiding in efficient data storage and feature identification for advanced analysis and predictive modeling.

1 Introduction

This project addresses the urgent need for advanced Electronic Health Record (EHR) data management by leveraging Natural Language Processing (NLP). Our primary goal is to convert unstructured EHR data into actionable formats through a meticulous approach, involving steps such as structured information extraction, keyword selection, and the use of regular expressions (regex).

To enhance healthcare providers' data access and management, we strategically implement classification models like Multinomial Naive Bayes, One-vs-Rest Logistic Regression, Word2Vec, and LightGBM for medical transcription categorization. Leveraging regular expressions and GPT-3.5, we extract medical entities for model training. The integration of the T5 model, known for its flexibility and encoder-decoder architecture, aims to transform unstructured transcriptions. Selected for its ease of access and extensive pre-training on diverse data, T5 ensures a robust understanding of intricate medical language patterns.

In summary, our project streamlines EHR data processing by extracting structured information and utilizing diverse classification models. The T5 model integration represents a key innovation, laying the groundwork for more efficient healthcare data management. The subsequent sections detail our methodologies, experiments, and results, providing insights into our approach and outcomes.

2 Related Work

The intersection of Natural Language Processing (NLP) and Artificial Intelligence (AI) in healthcare, especially Electronic Health Record (EHR) data management, has been extensively explored. Alanazi's research examines clinician perspectives on AI, identifying improvement areas and challenges. Complementing this, Mourad Sarrouti et al.'s study from Sumitovant Biopharma empirically compares encoder-only and encoder-decoder transformers, like T5, highlighting their effectiveness in biomedical text relation extraction.

Our project builds upon this foundation, addressing challenges and proposing a systematic methodology for converting unstructured EHR data into actionable formats. We leverage diverse classification models and integrate advanced AI models like T5 to enhance healthcare data management. The following sections provide detailed insights into our approach.

3 Approach

3.1 Extracting Structured Information from the Data

3.1.1 Regex based extraction:

The initial phase of the data extraction process involved the identification of key headings within medical transcripts by recognizing capitalized words. To enhance extraction efficiency, a dual approach was employed, comprising automated extraction of common capitalized headings and manual selection of keywords based on domain expertise. Subsequently, data extraction for each selected keyword was executed using regular expressions (regex) to capture pertinent information under each heading, ensuring precision in data retrieval.

3.1.2 GPT based extraction:

We also employed GPT-3.5 to extract medical entities from transcription texts complimenting the regex based extraction. Given the absence of gold standard labels, these methods served as a preliminary step aimed at identifying and categorizing entities such as Age, Disease, and Symptoms from the transcriptions.

3.2 Classification Model Development and Selection

In our project, we developed various classification models to systematically categorize medical transcriptions into distinct classes such as Gender, Medical Specialty, and Treatment Type. The selection of models was guided by the specific characteristics and requirements of the medical transcription data.

Data Transformation Technique: We employed Term Frequency-Inverse Document Frequency (TF-IDF) for transforming medical transcriptions. This technique effectively highlights important terms while minimizing the impact of frequent but less significant words, providing a balanced representation of the text data.

Implemented Models:

- **Multinomial Naive Bayes**: Chosen for its text data handling efficiency, particularly apt for categorizing medical transcriptions where simplicity and performance are key.
- **One-vs-Rest Logistic Regression**: Implemented to manage the multi-class nature of our data. It is particularly effective in addressing class imbalance issues in the dataset.
- **Word2Vec**: Selected for its ability to understand semantic context and relationships within medical transcription texts. This enhances the model's comprehension of medical terminology and nuances in language patterns.
- **LightGBM**: Opted for its scalability and speed in training on extensive medical transcription datasets. Its capabilities in handling categorical and text data, coupled with regularization and parallel computing, contribute to robust model performance and generalization.

3.3 Entity and Sequence Generation Model

Data Structuring: Following the entity extraction, we concatenated them into a unified column, thereby creating a structured format for the data. This consolidated column serves as the expected output, with the transcriptions constituting the input for our model. Data integrity and completeness were ensured by rigorously filtering the dataset, dropping rows where any of the values in the concatenated results column (Age, Diagnosis, Symptoms) were missing.

Transition to T5 Model: Initially, SciSpacy and Fine-tuned BioBERT, combined with LSTM, were explored for entity tagging and sequence generation tasks. While BioBERT demonstrated superior performance in entity tagging, it faced limitations in text generation, resulting in nonsensical output unsuitable for the project's requirements. Subsequently, the focus shifted to the T5 model. Various hyperparameters were experimented with to optimize its performance. The T5 model's inherent

flexibility and proficiency in handling sequence-to-sequence tasks proved to be more effective for the project's specific needs. This involved not only recognizing medical entities but also generating them in a structured sequence.

4 Experiments

4.1 Data

The Kaggle dataset, comprising 5,000 entries across six columns, presents a diverse array of medical transcripts spanning various specialties. Each entry encompasses essential details, including case descriptions, associated medical specialties, sample names, transcription text, and relevant medical terms. The dataset encapsulates a wide spectrum of medical scenarios, from allergic rhinitis presentations to bariatric consultations and cardiovascular examinations. This diversity offers a rich foundation for exploration, enabling the application of transfer learning techniques to extract meaningful patterns and insights from unstructured medical text.

4.2 Evaluation method

In evaluating our classification models, the primary metric selected is the Receiver Operating Characteristic Area Under the Curve (ROC AUC), chosen for its efficacy in handling class imbalances and providing a comprehensive assessment of model discrimination capabilities. This aligns with our focus on medical transcription data, where imbalances are prevalent, and accurate classification is paramount for healthcare decision-making. For assessing the performance of the T5 model in sequence generation, we utilize the ROUGE-L metric. This metric quantifies the overlap between the model's output and the gold standard, assessing both content accuracy and extraneous noise.

4.3 Experimental details

Table 1: Classification Models and Their Configurations

Model	Configurations
Naive Bayes	TfidfVectorizer (max_features=4500), alpha=1.0, fit prior=True
Logistic Regression	TfidfVectorizer (max_features=4500), C=1.0, penalty='l2', solver='lbfgs', max iter=100, class_weight='balanced'
LightGBM	num_estimators=120, learning_rate=0.01, max_depth=4, min_samples_leaf=2
Word2Vec	Vector size=100, window=5, min_count=1, sg=1 (skip-gram), workers=4

Table 2: T5 ROUGE-L Scores for Different Hyperparameter Settings

Learning Rate	Batch Size	Epochs	ROUGE-L		
			Precision	Recall	F1 Score
1e-4	4	3	0.2011	0.2734	0.2317
1e-4	4	10	0.2615	0.3459	0.2978
1e-4	1	10	0.1840	0.2571	0.2144
0.01	4	10	0.1524	0.2232	0.1811

4.4 Results

Table 3: Classification Task Results

Model	Average AUC ROC Score	Classification Task
Multinomial Naive Bayes	0.9092	Gender Classification
Logistic Regression	0.9693	Gender Classification
Multinomial Naive Bayes	0.7729	Specialty Classification
Logistic Regression	0.9461	Specialty Classification
Multinomial Naive Bayes	0.7541	Treatment Type Prediction
Logistic Regression	0.9441	Treatment Type Prediction
LightGBM	0.6245	Specialty Classification
Word2Vec	0.3783	Specialty Classification

The transformation process yielded a structured dataset, with medical information organized under relevant headings. The combined approach of automated and manual keyword selection ensures a comprehensive extraction of pertinent data.

Optimal results for the T5 model were achieved with the hyperparameters **learning rate of 1×10^{-4}, batch size 4 and 10 epochs.**

The ROUGE-L Scores for the above hyperparmeters are a **precision of 0.261, recall of 0.346 and F1 score of 0.298.**

5 Analysis

In our exploration of classification models—LightGBM, Word2Vec, One-vs-Rest Logistic Regression, and Multinomial Naive Bayes—we observed distinct performances influenced by various factors.

For LightGBM, its proficiency in medical transcription tasks arises from adeptly handling categorical features and text data. With an innate capability to manage medical specialties without intensive preprocessing, LightGBM's tree-based structure excels in capturing intricate relationships within medical text, explaining its strong overall performance. In contrast, Word2Vec's performance was suboptimal for medical transcription due to limited contextual understanding, especially in capturing the nuanced healthcare vocabulary. Trained on a non-medical corpus, Word2Vec struggled with out-of-vocabulary terms and contextual ambiguities in medical terminology, hampering its ability to provide meaningful representations.

Multinomial Naive Bayes achieved approximately 70% accuracy, showcasing simplicity, efficiency, and suitability for text classification. Its success stems from capturing essential patterns in the diverse medical transcription vocabulary, making effective predictions based on keyword prevalence and distinctive terms across different specialties. One-vs-Rest Logistic Regression excelled in classifying medical specialties owing to its compatibility with multi-class tasks. Its effectiveness in handling diverse and imbalanced data, simplicity, interpretability, and efficiency in capturing linear relationships contributed to its success in our context.

Transitioning to entity tagging and sequence generation, the T5 model's suboptimal performance in entity tagging can be attributed to its general pre-training, lacking extensive coverage of medical entity tagging. However, T5's architecture, designed for sequence generation tasks, led to excellent performance in this area, leveraging its training on diverse language tasks. We may have also been able to achieve better results had there been a larger corpus of data.

Exploring additional models, BioBERT, implemented with LSTM, demonstrated proficiency in entity tagging, benefitting from pre-training on medical data. In contrast, SciSpacy, employed for sequence generation, yielded suboptimal results, highlighting its limitations in specialized medical tasks.

6 Conclusion

In summary, our project addresses the critical need for structured electronic health record (EHR) data, presenting a systematic approach that transforms unstructured medical transcripts into actionable formats. The key achievement lies in the successful integration of diverse classification models, culminating in the incorporation of the T5 model for efficient sequence generation. This not only facilitates enhanced healthcare data management but also demonstrates improved classification performance, particularly in tasks related to gender, medical specialty, and treatment type prediction.

However, our work is not without limitations. The T5 model's suboptimal performance in entity tagging tasks reveals the need for domain-specific pre-training to better capture medical entities. Additionally, the Word2Vec model struggled with the nuanced healthcare vocabulary, emphasizing the importance of specialized pre-training for contextual understanding. Our work also acknowledges the challenges posed by class imbalances in medical transcription data, and while the selected metrics provide comprehensive evaluations, addressing these imbalances could further enhance model performance.

Looking ahead, we propose an ensemble approach that combines clinical BioBERT for entity tagging and T5 for sequence generation. This strategy aims to leverage clinical BioBERT's domain-specific knowledge for accurate entity tagging and T5's strengths in generating coherent sequences, potentially overcoming the limitations observed in individual models. As the field of natural language processing evolves, exploring advanced pre-training techniques and incorporating larger, domain-specific corpora could further enhance the robustness of our approach.

References

[1] Alanazi, Abdullah. *Clinicians' Views on Using Artificial Intelligence in Healthcare: Opportunities, Challenges, and Beyond.* Cureus, vol. 15, no. 9, e45255, 14 Sep. 2023. *doi:10.7759/cureus.45255*

[2] Borna, Sahar et al. *Artificial Intelligence Models in Health Information Exchange: A Systematic Review of Clinical Implications.* Healthcare (Basel, Switzerland), vol. 11, no. 18, p. 2584, 19 Sep. 2023. *doi:10.3390/healthcare11182584*

[3] Hossain, Elias et al. *Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review.* Computers in Biology and Medicine, vol. 155 (2023): 106649. *doi:10.1016/j.compbiomed.2023.106649*

[4] Hughes, Kevin S et al. *Natural language processing to facilitate breast cancer research and management.* The Breast Journal, vol. 26,1 (2020): 92-99. *doi:10.1111/tbj.13718*

[5] Crowson, Matthew G et al. *Towards Medical Billing Automation: NLP for Outpatient Clinician Note Classification.* medRxiv : The Preprint Server for Health Sciences, 2023.07.07.23292367. 12 Jul. 2023. *doi:10.1101/2023.07.07.23292367. Preprint*

[6] Wu, Yonghui et al. *Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network.* Studies in Health Technology and Informatics, vol. 216 (2015): 624-8.

[7] Hussain, Syed-Amad et al. *A natural language processing pipeline to synthesize patient-generated notes toward improving remote care and chronic disease management: a cystic fibrosis case study.* JAMIA Open, vol. 4,3 ooab084. 29 Sep. 2021. *doi:10.1093/jamiaopen/ooab084*

[8] Lineback, Christina M et al. *Prediction of 30-Day Readmission After Stroke Using Machine Learning and Natural Language Processing.* Frontiers in Neurology, vol. 12, 649521. 13 Jul. 2021. *doi:10.3389/fneur.2021.649521*

A Appendix

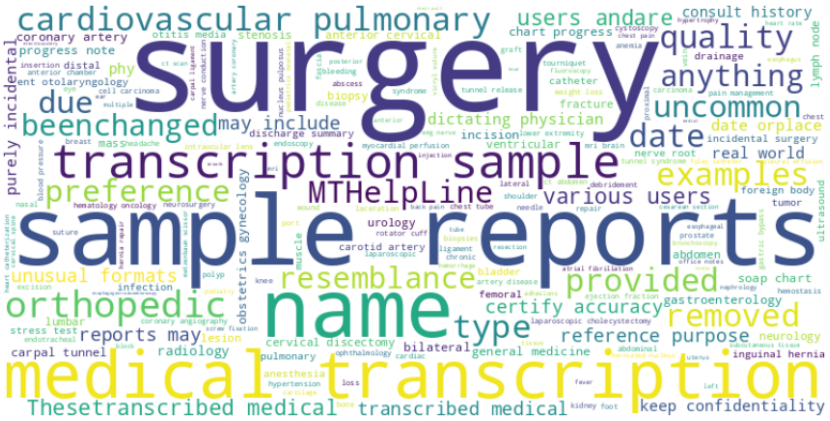

Figure 1: Word Cloud for the keyword column

Figure 2: Word Cloud for the transcription column

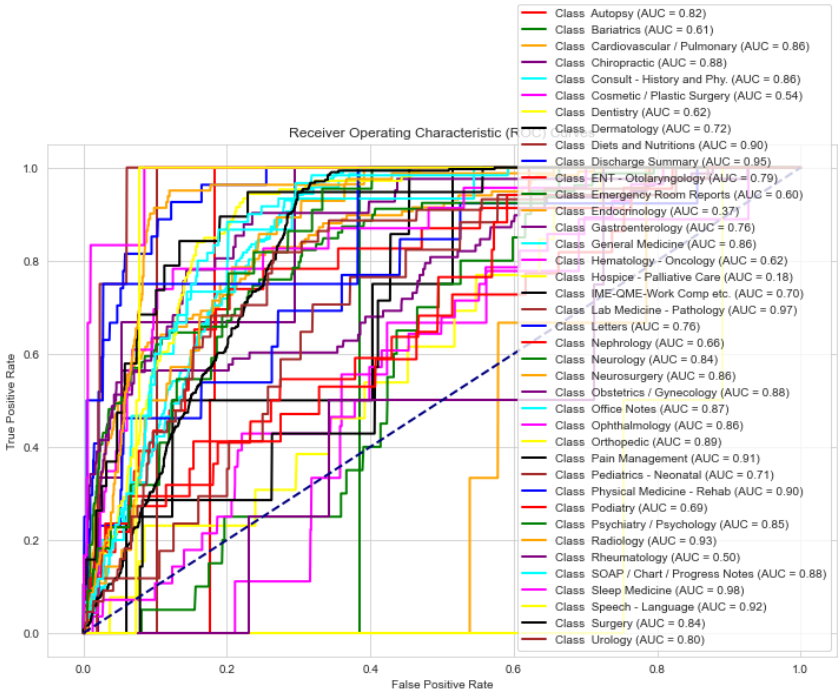

Figure 3: ROC AUC graph for speciality classification using Multinomial Naive base model

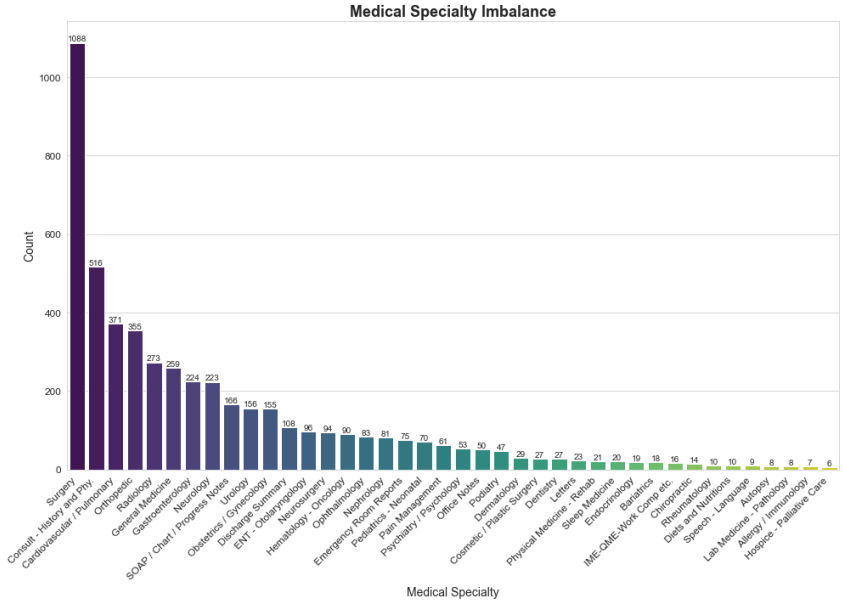

Figure 4: Medical Speciality imbalance

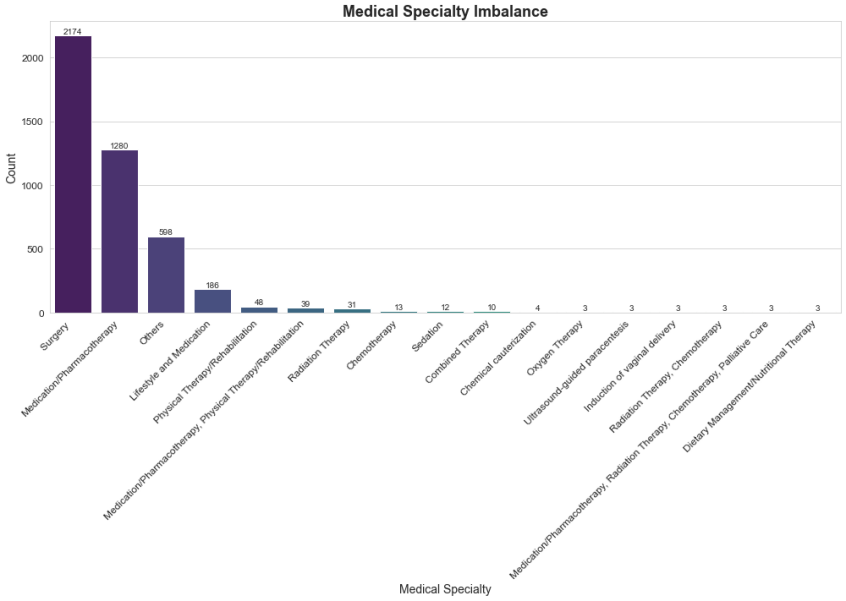

Figure 5: Medical Treatment imbalance