ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Classification of Clinical Texts via Sequential Forward Selection

[1]Mr. P S S K Sarma, [2]Satyawada Pavan Kumar, [3]Uppuluri Naga Satyanarayana,
[4]Penumatcha Sai Krishnam Raju,
[1]Associate Professor, Department of CSE, Rajamahendri Institute of Engineering & Technology, Bhoopalapatnam, Near Pidimgoyyi, Rajahmundry, E. G. Dist. A.P 533107.

[2,3,4]Student, Department of CSE, Rajamahendri Institute of Engineering & Technology, Bhoopalapatnam, Near Pidimgoyyi, Rajahmundry, E. G. Dist. A.P 533107.

Abstract

With far-reaching consequences for healthcare applications, clinical text categorization is an essential task in the field of Natural Language Processing. The main goal of this natural language processing research is to classify medical transcripts according to the underlying medical disorders. With the help of Sequential Forward Selection (SFS), a feature selection method carefully selected for its ability to reduce data dimension noise, we are able to achieve this. The goal of our study is to improve illness detection speed and accuracy by using SFS to boost pattern recognition efficiency and classification performance. The importance of Clinical Text Classification is highlighted in this study effort, which aims to optimize the process utilizing SFS.

Keywords— Clinical Text Classification, Medical transcripts, Sequential Forward Selection 33 1 2377 2358 Transcription title Sample medical transcriptions Keywords I.

INTRODUCTION

3848 x 1068 Important terms derived from voiceover Because of the immense amount of valuable information that may be extracted from unstructured clinical text data, clinical text categorization has emerged as a top priority in the healthcare industry. Because of its multidimensional significance and possible influence on healthcare administration and patient results, this area has grown in stature. Our proposed strategy for correctly classifying medical transcripts into their various specializations is part of our natural language processing effort. The features are text data taken from medical transcripts, and the target variable is the medical specialization. There are a number of critical phases to the project. Tokenization, stemming, and stop word removal are some of the pre-processing operations that may be necessary for the text data. Machine learning techniques like Logistic Regression, Support Vector Machines, and Categorical Boosting are trained and tested on the dataset after data preprocessing. Metrics like F1-score, recall, accuracy, and precision may be used to assess the performance of each model. At last, we may choose the most effective model and put it to use by assigning fresh medical transcripts to their correct medical specialty.

DATA DESCRIPTION

We scraped information from mtsamples.com to collect the Medical Transcriptions dataset from Kaggle.

TABLE I. DETAILED DATASET DESCRIPTION

Column Names	Missing Values	Missing Value %	Unique Values	Column Definition
Description	0	0	2348	Short description of transcription
medical specialty	0	0	40	Medical specialty classification of transcription
sample name	0	0	2377	Transcription title
Transcription	33	1	2358	Sample medical transcriptions
Keywords	1068	21	3848	Relevant keywords from transcription

METHODOLOGY

A systematic method for preparing data and building models is critical in the field of Natural Language Processing (NLP) projects. Researchers and practitioners alike will find this part helpful as it lays out the fundamentals of data preparation for analysis. In the first stage, known as data cleaning, an exhaustive evaluation of the dataset is carried out.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

The goal of this procedure is to fix problems like inconsistent formatting, missing data, or duplication. Thorough cleaning guarantees the data's dependability, which is vital for future analysis. After data cleaning, the next step is preprocessing, which is all about getting the raw text data ready for analysis. Lemmatization, POS (Part-of-Speech) tagging, and tokenization are tasks that fall under this level. By following these procedures, raw text may be transformed into a structured and analytically-ready format. As part of getting the data ready for the model, the dataset is split into three separate subsets: training, validation, and testing. Additionally, it requires transforming textual information into a numerical form that is suitable with ML models. When training and evaluating a model, this stage is vital.

The selection of the most important characteristics from the dataset is known as feature selection, and it is an essential operation. The foundation of machine learning models are these qualities. Feature selection techniques might vary from choosing the most common terms to using statistical approaches to find predictive characteristics. By zeroing in on relevant data, this phase improves model efficiency. Training and evaluating machine learning models are at the heart of the last stage, model creation. Depending on the goals of the study, this step may include either a classification model or a regression model. The success of natural language processing (NLP) initiatives depends on the careful implementation of certain methodological procedures. To answer the research questions or achieve the goals, they ensure that the data used is accurate and reliable and that the machine learning models that are created are effective.

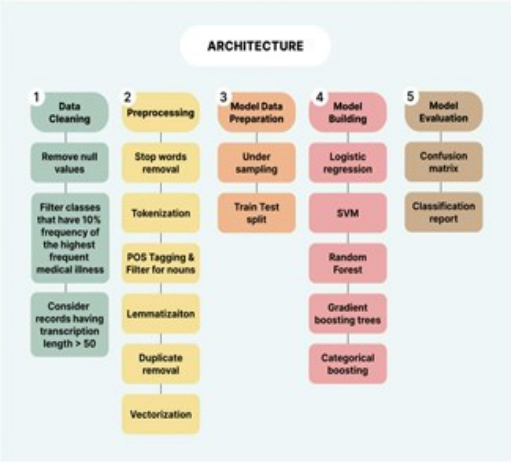

Fig. 1. Detailed Workflow of the Model

HIGH LEVEL SOLUTION FLOW

Subject: Data Cleaning Data cleaning is a critical part of natural language processing (NLP) projects since it guarantees that the data used for analysis is correct and trustworthy. A prevalent problem with datasets is the existence of null values, which may impact the efficiency of natural language processing algorithms. Eliminating the null values from the dataset will solve this problem. To do this, we must first find the rows or columns that have null values and then remove them. After removing the null values, the dataset is further processed by removing classes with less than 10% of the maximum entries for that class. The records are then sorted by selecting only those with transcription lengths greater than 50. B. Preprocessing 1) Eliminating Stop Words from the Nltk Library and Specifically for the Domain: Common words that are routinely eliminated from text during natural language processing preprocessing are called stop words. These words are thought to not significantly affect the overall meaning of a phrase. You may get a list of frequently used stop words in text data in the Natural Language Toolkit (nltk) package. To further enhance the performance of natural language processing models, it is also possible to add or delete domain-specific stop words from the text input. In this example, we will use medical stop words.

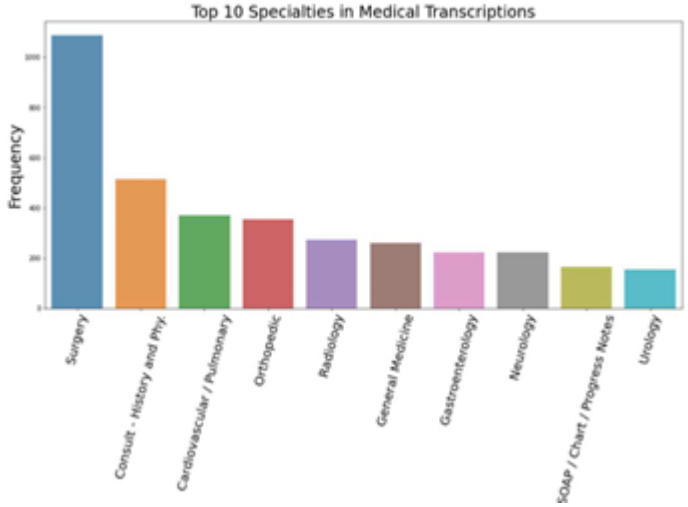

Fig. 2. Ten most Frequent Medical Illnesses in the Dataset

The nltk library is used for tokenization, which is the act of dividing a phrase or text into smaller parts called tokens. Word tokenize and sent tokenize are two of the many tokenization techniques provided by the nltk library. Word tokenize breaks text into words while sent tokenize breaks text into sentences.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

3. Using the nltk library for POS tagging and noun filtering: Classifying sentences according to their constituent elements of speech is what it's all about. Use the POS tagging capabilities provided by the nltk library to filter for nouns and other specified parts of speech. If you want to find the most important terms in a book to study further, this could help. 4) WordNetLemmatizer for Lemmatization: Lemmatization decreases the dimensionality of text data and enhances the performance of natural language processing models by reducing words to their base or root form. You may apply lemmatization to text data using the WordNetLemmatizer that is part of the nltk package. 5) Eliminate duplicate features from the feature list: Overfitting and poor model performance might result from feature duplication. Hence, it's critical to eliminate characteristics that are duplicates. Several techniques exist for doing this, such as the set() function, which converts the list to a set and then back to a list again, or the pandas drop duplicates() function. 6) TfidfVectorizer for vectorization: This is a method for transforming text input into numerical format that ML models use. Using the term frequency inverse document frequency (TF-IDF) approach, which prioritizes less frequent terms in the text data, the TfidfVectorizer can vectorize text data. That way, natural language processing models can function better.

Fig. 3. Word Cloud for Medical Transcripts

Class C: Preparing Data for Models 1) Train test split: We divided the data in half, using half for practice and half for the test. This ensures that the model isn't only learning the practice set but also learns generalizable patterns by allowing us to see how well it will do on fresh, unseen data. 2) Random UnderSampler for Under Sampling: When one class is much more common than the others in a dataset, a problem known as class imbalance arises. One way to fix this is by using under sampling, a method used in machine learning. Biased models that fail to adequately represent minority groups may result from this. The imbalanced-learn Python library provides a function called RandomUnderSampler that may be used to systematically reduce samples from the class with the largest number of members until the distribution of the classes is more evenly distributed. The model's accuracy when applied to the minority class may be enhanced in this way. Chapter D. Forward Feature Selection for Feature Selection The term refers to the steps used to determine which dataset attributes are most relevant for usage in a machine learning model. Reducing the dataset's dimensionality, eliminating noise and unnecessary features, and making the model more interpretable are all ways this might boost model performance. An approach to feature selection known as "forward feature selection" adds features to a model iteratively depending on how they enhance its performance, rather than beginning with a full set of features. The first step is to train a model using each feature separately and then choose the one that performs the best. Next, we iterate by training models with each conceivable combination of the characteristics that were previously chosen, and we choose the one that performs the best. This cycle continues until either the performance target or the specified number of features is met. One of the many benefits of forward feature selection is that it simplifies the process of identifying the most significant characteristics in a dataset while still being computationally efficient. Nevertheless, overfitting might occur if the dataset is too small in comparison to the number of features used. Hence, cross-validation is a must for making sure the chosen features are good at applying to new data.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

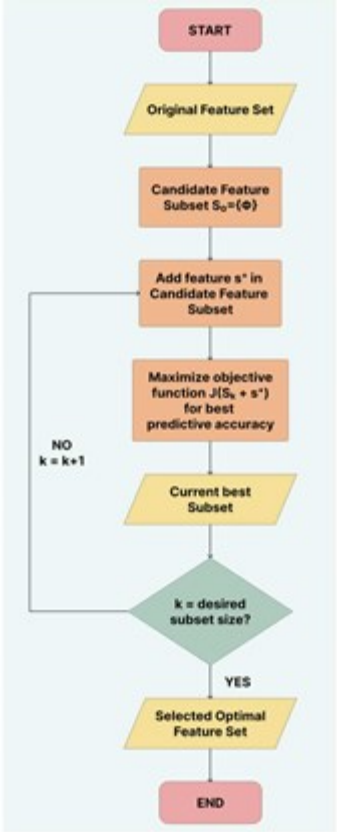

Fig. 4. Flowchart for Sequential Forward Selection (Feature Extraction Algorithm)

Part E: Creating Models 1) One statistical model that aims to estimate the likelihood of an outcome is logistic regression. The connection between a group of independent variables is what makes it tick. When dealing with difficulties involving several classes, logistic regression may be a useful tool. SVM is a supervised learning method that is used for regression analysis and classification. Support vector machines (SVMs) sort data into categories by locating a hyperplane in a three-dimensional space that does this job optimally. Because they use non-linear kernel functions to convert the input data into higher-dimensional feature spaces, support vector machines (SVMs) are able to handle data that is not linearly separable. Thirdly, Random Forest is a well-liked ensemble learning approach for feature selection, regression, and classification. During training, it builds a forest of decision trees and then uses the mean prediction from each tree to produce a class. Random Forest is able to process a high volume of input features while being resilient in the face of outliers and noisy data. 4) Gradient Boosting trees: This boosting approach repeatedly trains each model to fix the mistakes of the previous one, combining several weak models into a strong one. An enhancement of Gradient Boosting, Gradient Boosting trees employ decision trees as its weak models. The approach incorporates decision trees into the model in an iterative fashion, with each tree being trained using the residual mistakes from the ones before it. Because of its famed capability to manage complicated and non-linear interactions between the input characteristics and the target variable, Gradient Boosting trees find utility in both regression and classification problems. Fifthly, there's categorical boosting, which is a kind of gradient boosting made for classification data. Use of decision trees with categorical splits rather than continuous splits and the incorporation of categorical encoding methods like target encoding and one-hot encoding are the building blocks of categorical boosting. One of categorical boosting's well-known uses is in classification problems; it excels at dealing with unbalanced datasets and features with high cardiacity.

EXPERIMENTATION

• We developed a way to selectively extract nouns from transcripts using POS tagging as part of our data processing experiments aimed at reducing the data set for easier feature selection. • To reach specific conclusions, we documented the outcomes of experiments with varying numbers of courses. • We tried out various feature counts throughout the feature selection process; for example, we found that 10 features was underfitting and 20 features was overfitting; so, we settled on 15 features. • Afterwards, we tested many models to see which one was most effective for this problem statement. These models included Logistic Regres-sion, SVM, Random Forest, Gradient Boosting trees, and Categorical Boosting. The majority of the transcriptions were found to be same when mapped to various specializations, according to our Cosine and Jaccard similarity analyses.

RESULTS

Here you should lay down the project's findings, including the insights you obtained and how they connected to the initial business issue. • We said at the end of the research that there is a severe lack of data and that a larger dataset is necessary for accurate prediction. • Using Categorical Boosting, we were able to get a 99% accuracy rate for the two medical

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

conditions: "Surgery" and "Consultation and History."

.

	precision	recall	f1-score	sup
Surgery	0.98	1.00	0.99	
Consult - History and Phy.	1.00	0.98	0.99	
accuracy			0.99	
macro avg	0.99	0.99	0.99	
weighted avg	0.99	0.99	0.99	

Fig. 5. Classification Report for 2 Medical Illnesses using Categorical Boosting

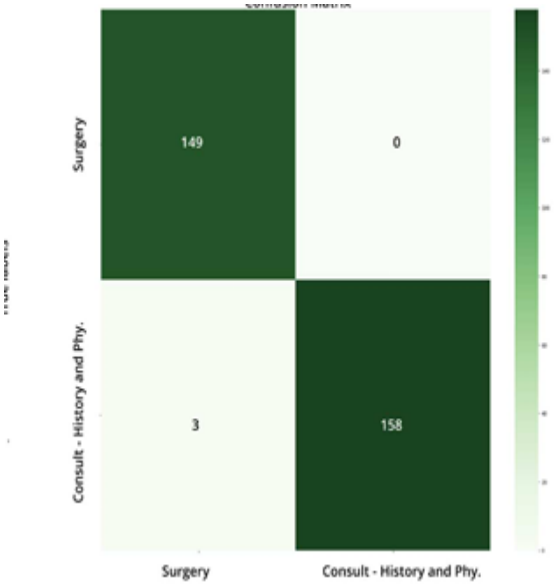

Fig. 6. Confusion Matrix (Heatmap)

when we use categorical boosting for two medical diseases, the accuracy is 75%. However, when we use categorical boosting for three medical illnesses, namely "surgery," "consultation and history," and "cardiovascular/pulmonary," the accuracy drops to 72%. Misclassification occurs when records are being assigned to the Surgery and Consultation and History courses whereas, in fact, many of the transcripts from other classes describe surgeries related to medical specialties or patient histories. The cardiovascular and pulmonary classes, for instance, include transcripts that detail heart surgeries or make reference to a patient's cardiac history.

.

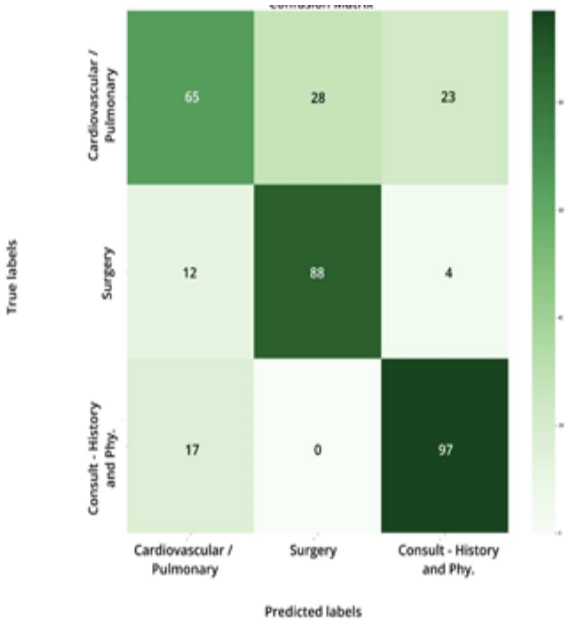

Fig. 7. Confusion Matrix (Heatmap) for 3 Medical Illnesses using Categorical Boosting

	precision	recall	f1-score	support
Cardiovascular / Pulmonary	0.69	0.56	0.62	116
Surgery	0.76	0.85	0.80	104
Consult - History and Phy.	0.78	0.85	0.82	114
accuracy			0.75	334
macro avg	0.74	0.75	0.74	334
weighted avg	0.74	0.75	0.74	334

Fig. 8. Classification Report for 3 Medical Illnesses using Categorical Boosting

CONCLUSION

We may limit the number of categories we need to explore by using our subject matter expertise to group comparable categories together. While hand-crafted features may improve the dataset's performance, they may not be applicable to other transcription datasets. In order to properly assign the transcriptions to their corresponding medical categories, we have determined that more data is necessary. In order to do multiclass classification, future work may need string splitting. According to our findings, more information is required for reliable medical transcribing classification. Our ability to get the requisite precision has been hindered by the existing dataset's modest size. We want to remedy this by

breaking the transcriptions into more manageable chunks in the future. Because of this, we may use a multiclass classification strategy to classify the medical information with more precision and subtlety. Our medical transcribing system will be more effective and trustworthy as a consequence of this, as we anticipate that the accuracy of our categorization findings will be much enhanced. To put it more simply, we need to gather more data and then break it down into more manageable chunks. Because of this, we will be able to better sort the transcriptions into their respective medical fields.

REFERENCES

[1]. Yao, L., Mao, C. & Luo, Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Mak 19 (Suppl 3), 71 (2019).

[2]. Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. Deep Learning for Health Informatics. IEEE journal of biomedical and health informatics, 21(1), 4–21, (2017).

[3]. Rao, S. Govinda, R. Rambabu, and P. VaraPrasada Rao. "Modified Hierarchical Clustering algorithms to Evaluate the Similarities of Growth Factor IR Inhibitors by Using Regression Analysis." In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pp. 1-8. IEEE, 2018.

[4]. A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. G. Cortina Januchs and D. Andina. Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, Glendale, AZ, USA, pp. 2845-2850, (2010).

[5]. Garla, V., Taylor, C., & Brandt, C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. Journal of biomedical informatics, 46(5), 869–875, (2013).

[6]. Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. What can natural language processing do for clinical decision support?. Journal of biomedical informatics, 42(5), 760–772, (2009).

[7]. Özlem Uzuner, Ira Goldstein, Yuan Luo, Isaac Kohane. Identifying Patient Smoking Status from Medical Discharge Records. Journal of the American Medical Informatics Association, Volume 15, Issue 1, Pages 14–24, January 2008.

[8]. Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. Medical Semantic Similarity with a Neural Language Model. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14). Association for Computing Machinery, New York, NY, USA, 1819 1822, (2014).

[9]. Last, M., Kandel, A., & Maimon, O. Information-theoretic algorithm for feature selection. Pattern Recognition Letters, 22(6-7), 799-811, (2001).

[10]. Kudo, Mineichi & Sklansky, Jack. Sklansky, J.: Comparison of Algorithms that Select Features for Pattern Classifiers. Pattern Recognition 33, 25-41. Pattern Recognition. 33. 25-41. (2000).

[11]. D. Xiao and J. Zhang. Importance Degree of Features and Feature Selection. 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China, pp. 197-201. (2009)

[12]. Agrawal, K. K. ., P. . Sharma, G. . Kaur, S. . Keswani, R. . Rambabu, S. K. . Behra, K. . Tolani, and N. S. . Bhati. “Deep Learning-Enabled Image Segmentation for Precise Retinopathy Diagnosis”. *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 12s, Jan. 2024, pp. 567-74, https://ijisae.org/index.php/IJISAE/article/view/4541

[13]. Samota, H. ., Sharma, S. ., Khan, H. ., Malathy, M. ., Singh, G. ., Surjeet, S. and Rambabu, R. . (2024) “A Novel Approach to Predicting Personality Behaviour from Social Media Data Using Deep Learning”, *International Journal of Intelligent Systems and Applications in Engineering*, 12(15s), pp. 539–547. Available at: https://ijisae.org/index.php/IJISAE/article/view/4788