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.ijsem.org
Vol 19, Issue.1 March 2025

CLASSIFYING CLINICAL TEXT THROUGH SEQUENTIAL FORWARD SELECTION

G.Lalitha [1], Dr.N.Deepak Kumar [2],
PG Scholar [1], Dept of CSE, Sree Rama Engineering College, Tirupati – 517507.
Professor [2], Dept of CSE, Sree Rama Engineering College, Tirupati – 517507.

Abstract—

A crucial task in the field of Natural Language Processing, clinical text categorization has important consequences for healthcare applications. Classifying medical records into their respective diseases is the main goal of this natural language processing study. Our feature selection approach of choice, Sequential Forward Selection (SFS), was hand-picked for its ability to reduce data dimension noise, so we may use it to our advantage. Our study aims to improve classification performance and pattern recognition efficiency using SFS, which will lead to faster and more accurate illness detection. With an emphasis on improving the process utilizing SFS, this study work highlights the critical importance of Clinical Text Classification. Subjects—Medical Transcripts, Clinical Text Classification, Sequential Forward Selection

INTRODUCTION

The need to get into the quantity of information present in unstructured clinical text data has elevated clinical text categorization to the status of a crucial task within the healthcare area. The potential influence on healthcare administration and patient outcomes, as well as the field's multidimensional significance, have contributed to its rise to prominence.

We provide a strategy for correctly categorizing medical transcripts into their respective specializations in this natural language processing research. In this context, the medical specialty serves as the target variable, while the characteristics are the text data retrieved from the medical transcripts. There are a number of critical phases to the project.

Some of the pre-processing steps that may be necessary for the text data include stemming, tokenization, and stop word removal. Machine learning techniques like Logistic Regression, Support Vector Machines, and Categorical Boosting are trained and tested on pre-processed datasets. Metrics like F1-score, recall, accuracy, and precision may be used to assess the performance of each model. The last step is to choose the most effective model and use it to categorize newly uploaded medical transcripts into their corresponding medical subspecialties.

DATA DESCRIPTION

The Medical Transcriptions dataset was obtained from mtsamples.com through content scraping and then imported into Kaggle.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN:2454-9940
www.ijsem.org
Vol 19, Issue.1 March 2025

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

Data preparation and model building in the field of Natural Language Processing (NLP) projects must adhere to an organized strategy. This section provides academics and practitioners with a clear framework by outlining the major phases in data preparation for analysis. The first step is data cleansing, which entails a thorough evaluation of the dataset. Problems like missing data, duplication, or inconsistent formatting are what this method is trying to fix. Thorough cleaning guarantees the data's dependability, which is essential for future analysis.
After data cleaning, the next step is preprocessing, which is all about getting the raw text data ready for analysis.

Lemmatization, tokenization, and POS (Part-of-Speech) tagging are all tasks that fall under this level. By following these procedures, raw text may be transformed into a structured and analytically-ready format.

Part of getting the data set ready for a model is splitting it up into different portions for things like training, validation, and testing. In addition, it takes textual input and transforms it into a numerical representation that machine learning models can understand. Training and evaluating the model depend on this stage.

A crucial step in data mining is feature selection, which entails picking out the most important properties from the information.

Machine learning models are built on top of these properties. Some feature selection techniques include using statistical approaches to find predictive traits, while others include choosing the most common terms. Refining the model's emphasis to relevant data improves its efficiency.

Training and evaluating machine learning models are at the heart of the last stage, model creation. During this stage, many kinds of models are considered, such as regression or classification models, based on the goals of the study.
The success of natural language processing (NLP) initiatives depends on the careful implementation of these methodological procedures. They ensure that the data used is accurate and reliable, and that the machine learning models that are created are effective in answering the research questions or accomplishing the goals.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN:2454-9940
www.ijsem.org
Vol 19, Issue.1 March 2025

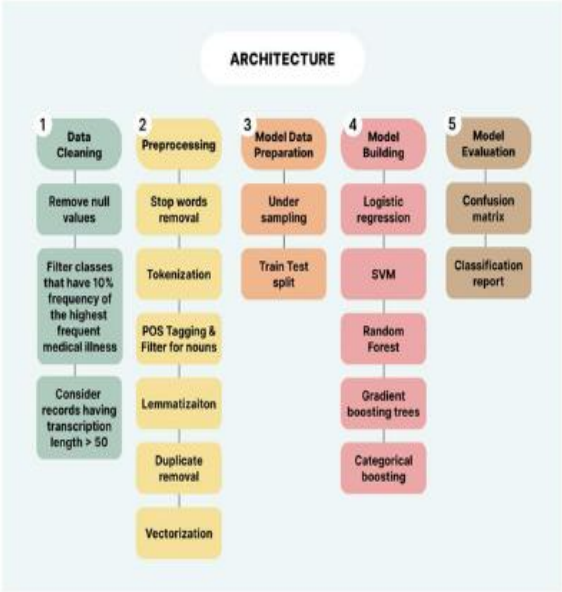

Fig. 1. Detailed Workflow of the Model

HIGH LEVEL SOLUTION FLOW

Data Cleaning

To guarantee correct and trustworthy data for analysis, data cleaning is a crucial stage in natural language processing projects. 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. The process involves finding the rows or columns that have null values and removing 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. In a subsequent step, the records are sorted by selecting only those with transcription lengths greater than 50.

Preprocessing

Disable the removal of terms from the nltk library and any domain-specific ones: During natural language processing (NLP) preprocessing, it is standard practice to eliminate terms that are often used but do not contribute significantly to the meaning of the phrase. To help with text data cleaning, the Natural Language Toolkit (nltk) package has a list of frequently used stop words. Furthermore, in order to enhance the functionality of natural language processing models, it is possible to construct and exclude domain-specific (here, medical) stop words from the text input.

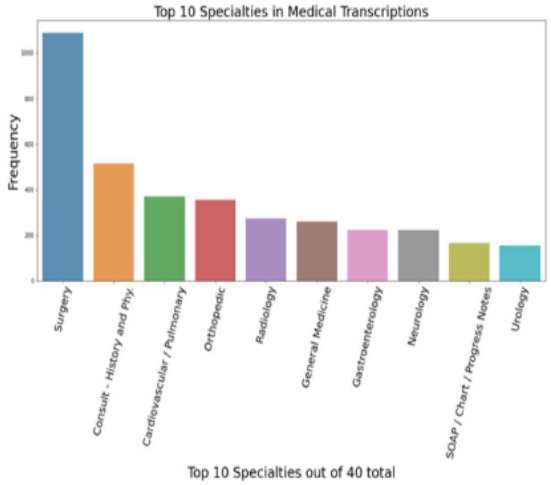

Fig. 2. Ten most Frequent Medical Illnesses in the Dataset

Tokenization using the nltk library is the second method. Tokenization involves dividing a text or phrase into smaller pieces, or 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. 3. Using the nltk library for POS tagging and noun filtering: Recognizing and classifying the many components of a phrase is what it is all about. If you want to filter for nouns or other particular parts of speech, you may utilize the POS tagging capabilities provided by the nltk library. Finding the most important words in a text to analyze further may be facilitated by this. 4) WordNetLemmatizer for Lemmatization: Lemmatization is a technique for lowering a word to its fundamental or root form; it may

ISSN:2454-9940
www.ijsem.org
Vol 19, Issue.1 March 2025

help reduce the dimensionality of text data and enhance the efficiency of natural language processing models. To apply lemmatization to text data, you may make use of the WordNetLemmatizer that is part of the nltk package. In order to prevent overfitting and poor model performance, it is important to remove duplicate features from the feature list. Hence, it's critical to eliminate features that are duplicated. There are a number of ways to do this, such as converting the list to a set and back to a list again using the set() function or using the drop duplicates() function in pandas. 6) Using TfidfVectorizer for vectorization: Transforming text input into numerical form for use in ML models is known as vectorization. Using the term frequency inverse document frequency (TF-IDF) approach, which prioritizes less frequent terms in the text data, the TfidfVectorizer may conduct vectorization on text data. Natural language processing models may benefit from this.

Fig. 3. Word Cloud for Medical Transcripts

Model Data Preparation

Train test split: We divided the data in half, using half for practice and half for the test. To ensure the model isn't only learning the practice set but really discovering generalizable patterns, this allows us to see how well it would do on fresh, unseen data.

Under sampling using

RandomUnderSampler: Under sampling is a machine learning approach that helps deal with class imbalance, which happens when one class is much more common than the others in a dataset. Because of this, biased models may be developed that have a negative impact on minority groups.

To make the class distribution more balanced, you may use RandomUnderSampler, a function supplied by the imbalanced-learn Python package, to randomly remove samples from the dominant class. The model's accuracy when applied to the minority class can be enhanced in this way.

The D. Forward Feature Selection Method for Feature Selection

To train a machine learning model, feature extraction involves sorting through a dataset and picking out the most relevant characteristics. Reducing the dataset's dimensionality, eliminating noise and unnecessary features, and making the model more interpretable are all ways this might boost model performance. Using an empty collection of features as a starting point, forward feature selection gradually adds features depending on the increase in model performance. The first step is to train a model using each feature separately and then choose the one that performs the best. After then, it's back to the drawing board to see which feature combination yields the best performance by training models with every conceivable combination of the features that were previously chosen. This procedure is carried out until either the required performance

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN:2454-9940
www.ijsem.org
Vol 19, Issue.1 March 2025

level or the specified number of features is attained.

Being computationally efficient and offering a straightforward technique to discover the most significant characteristics in a dataset are two of the many benefits of forward feature selection. But if there are too many features chosen in relation to the dataset size, overfitting might occur.

Therefore, it is vital to do cross-validation to ensure that the chosen characteristics generalize well to unseen data.

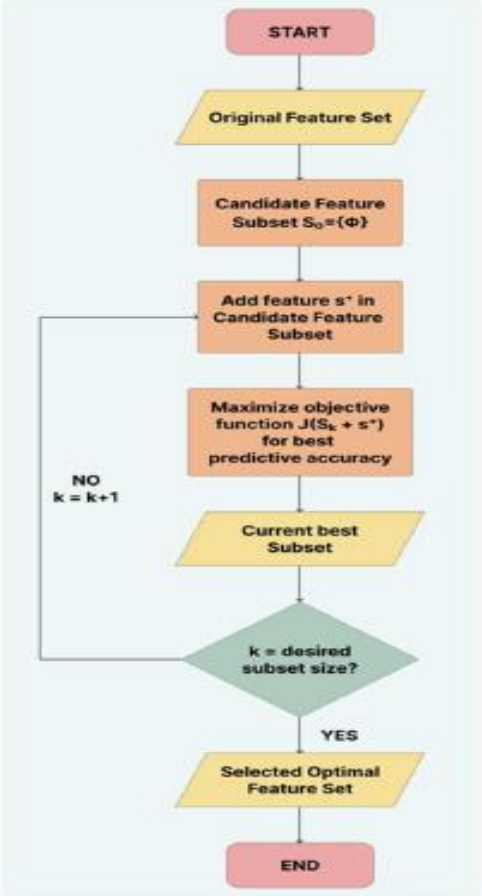

Fig. 4. Flowchart for Sequential Forward Selection (Feature Extraction Algorithm)

Model Building

One statistical model that aims to estimate the likelihood of an outcome is logistic regression. A collection of independent variables is used to represent the connection, which is how it works. Problems involving several classes in a classification model may also be addressed using logistic regression.

Second, support vector machines (SVMs) are a subset of supervised learning algorithms that find usage in regression and classification studies. SVMs function by locating a hyperplane in a space with a high degree of dimensionality that effectively divides the data points into distinct groups. By using non-linear kernel functions, SVMs convert the input data into feature spaces with greater dimensions, allowing them to handle data that is not linearly separable.

Random Forest: One of the most well-known ensemble learning methods, it is used for feature selection, regression, and classification. The method relies on training a large number of decision trees and then producing a class that represents the average or mode of those trees' predictions. Despite dealing with noisy data and outliers, Random Forest is able to process a high volume of input information.

Gradient Boosting trees: This boosting approach repeatedly trains each model to fix the mistakes of the original model, combining several weak models into a strong one. One variant of Gradient Boosting that makes use of decision trees as its weak model is Gradient Boosting trees. To train the model, the method iteratively incorporates decision trees, with each tree learning from the mistakes left by the ones before it. The capacity to manage complicated and non-linear interactions between the input characteristics and the goal variable is what makes Gradient Boosting trees so useful for classification and regression applications. 5) Categorical Boosting: This technique is a

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN:2454-9940
www.ijsem.org
Vol 19, Issue.1 March 2025

variant of Gradient Boosting and was developed with categorical data in mind. Rather of employing continuous splits, decision trees in categorical boosting use categorical splits, and categorical encoding methods like one-hot and target encoding are used. The capacity to handle unbalanced datasets and high-cardinality categorical features has made categorical boosting a popular choice for classification problems.

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. In order to reach specific conclusions, we conducted experiments with varying numbers of classes and documented the outcomes.

• We thought of using 15 features after experimenting with various numbers of features; for example, 10 features was underfitting and 20 features was overfitting.

• 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.

• Using Cosine and Jaccard similarity, we found that many transcriptions were identical and assigned to distinct fields of study.

RESULTS

The project's findings, including any insights obtained and their relevance to the initial business challenge, should be presented in this section.

• We said at the end of the research that there is a severe lack of data and that more records are needed to make accurate projections.

• Using Categorical Boosting, we were able to get a 99% accuracy rate for the two medical conditions: "Surgery" and "Consultation and History."

	precision	recall	f1-score	support
Surgery	0.98	1.00	0.99	149
Consult - History and Phy.	1.00	0.98	0.99	161
accuracy			0.99	310
macro avg	0.99	0.99	0.99	310
weighted avg	0.99	0.99	0.99	310

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

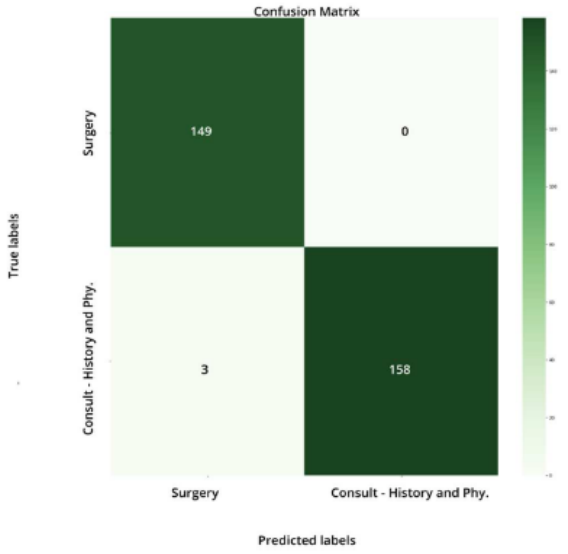

Fig. 6. Confusion Matrix (Heatmap) for 2 Medical Illnesses using Categorical Boosting

Category Boosting achieves an accuracy of 75% when three medical conditions—"Surgery," "Consultation and History," and "Cardiovascular/Pulmonary"—are selected. Due to the fact that many transcripts from other courses describe surgeries pertaining to certain medical specialties or patient histories, the data are being incorrectly mapped to the Surgery and Consultation and History classes. As an example, transcripts describing heart surgeries or discussing patients' cardiac histories are

ISSN:2454-9940
www.ijsem.org
Vol 19, Issue.1 March 2025

included in the cardiovascular and pulmonary target classes.

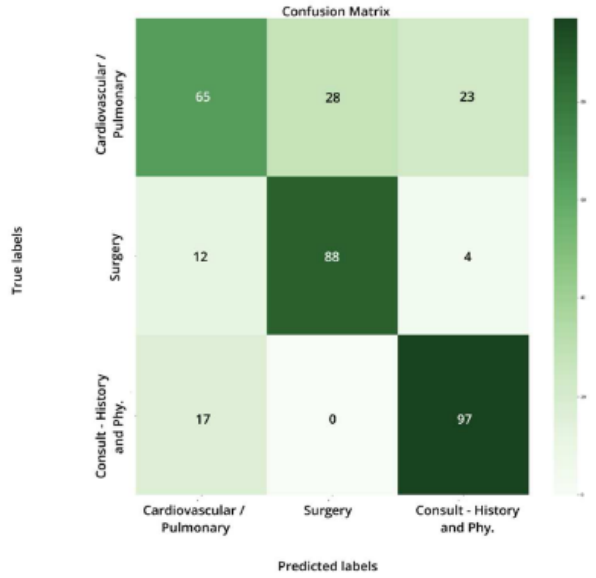

Fig. 7. Confusion Matrix (Heatmap) for 3 Medical Illnesses using Categorical Boosting

	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 overall number of categories we need to evaluate by grouping comparable ones together, using our subject matter expertise. Although manually creating features could improve the dataset's performance, these characteristics might not be applicable to other transcription datasets. We have determined that further data is needed to properly categorize the transcriptions according to their medical meaning. In order to do multiclass classification, future work may need string splitting. More data is required to properly categorize transcriptions into medical terms, according to our findings. We have been unable to get the required level of accuracy due to the present dataset's inadequacy. We want to solve this by breaking the transcriptions down 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 depth. We anticipate this will greatly enhance the precision of our categorization outcomes, leading to a more practical and dependable medical transcribing system. More data, broken down into smaller, more particular chunks, is what we need, to put it simply. 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] 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).

[4] Garla, V., Taylor, C., & Brandt, C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management.

ISSN:2454-9940
www.ijsem.org
Vol 19, Issue.1 March 2025

Journal of biomedical informatics, 46(5), 869–875, (2013).

[5] 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).

[6] 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.

[7] 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).

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

[9] 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).

[10] 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)