ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Applying Sequential Forward Selection for Clinical Text Classification

[1] S. Gouthami, [2] M. Jhansi

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Article Info

Received: 30-04-2025 Revised: 16-06-2025 Accepted: 28-06-2025

Abstract—

Classifying clinical texts is an important task in the field of Natural Language Processing with far-reaching consequences for healthcare applications. Classifying medical transcripts into their respective medical conditions is our main goal in this natural language processing study. The strength of Sequential Forward Selection (SFS), a feature selection approach carefully selected for its ability to simplify data dimensions, is harnessed to do this. Our study aims to improve classification performance and pattern recognition efficiency using SFS, which will lead to faster and more accurate illness detection. 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

INTRODUCTION

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. The potential influence on healthcare administration and patient outcomes, as well as the field's multidimensional significance, has contributed to its rise to prominence. Our proposed strategy for correctly classifying medical transcripts into their various specializations is part of our natural language processing effort. Features are text data taken from medical transcripts, and the goal variable is the medical specialization. There are a number of critical phases to the project. Preprocessing the text data may include operations like stemming, tokenization, and stop word removal, among others. Following data preprocessing, the dataset is used to train and assess a variety of machine learning techniques, including logistic regression, support vector machines, and categorical boosting, among others. Several measures may be used to assess each model's performance, including F1-score, recall, accuracy, and precision. Last but not least, fresh medical transcripts may be classified into their respective medical specialties using the most effective model.

DATA DESCRIPTION

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

Table 1 DETAILED DATASET DESCRIPTION

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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 essential stages required in data preparation for analysis. In the first stage, known as data cleaning, an exhaustive evaluation of the dataset is carried out. Problems like missing data, duplication, or inconsistent formatting are what this method is trying to fix. Data dependability depends on its integrity, which is why thorough cleansing is so important. Data cleaning is the first step in the preprocessing phase, which aims to prepare raw text data for analysis. At this level, tasks including lemmatization, POS (Part-of-Speech) tagging, and tokenization are included. 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. In addition, it requires transforming textual information into a numerical form that can be used by machine learning algorithms. In order to train and evaluate the model, this phase is vital. Finding the most important characteristics in the dataset is an essential part of the feature selection process. The foundation of machine learning models is these qualities. Some feature selection techniques include using statistical approaches to find predictive traits, while others include picking the most common terms. By zeroing in on relevant data, this phase improves model efficiency. Model construction, the last stage, is all about training and evaluating ML models. During this stage, many kinds of models are considered, such as regression or classification models, based on the goals of the study. In conclusion, natural language processing (NLP) initiatives cannot be successful without carefully following these methodological stages. They ensure that the data used is accurate and reliable, and that the machine learning models that come out of it are effective in answering the research questions or accomplishing the goals.

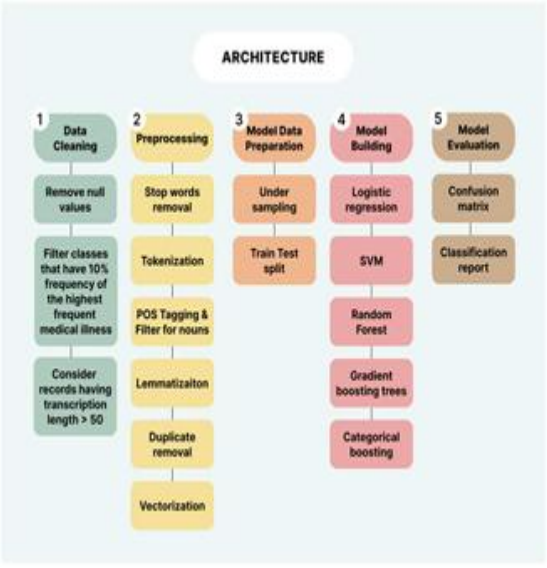

Fig. 1. Detailed Workflow of the Model

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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. The existence of null values is a prevalent problem in datasets and may impact the effectiveness of natural language processing algorithms. Eliminating the null values from the dataset will solve this problem. The rows or columns with null values have to be found and removed or updated accordingly. Classes with less than 10% of the maximum entries for that class are filtered out of the dataset once the null values have been eliminated. In a subsequent step, we narrow down the records to those with transcription lengths greater than 50. B. Preprocessing 1) Eliminating the following phrases 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. You may get a list of frequently used stop words in text data in the Natural Language Toolkit (nltk) package. It is also possible to enhance the performance of natural language processing models by adding or removing domain-specific stop words from the text input; in this example, medical stop words.

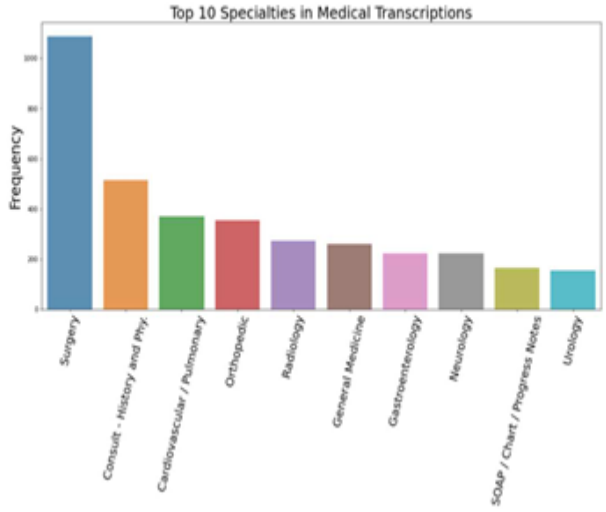

Fig. 2. Ten most Frequent Medical Illnesses in the Dataset

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. The existence of null values is a prevalent problem in datasets and may impact the effectiveness of natural language processing algorithms. Eliminating the null values from the dataset will solve this problem. The rows or columns with null values have to be found and removed or updated accordingly. Classes with less than 10% of the maximum entries for that class are filtered out of the dataset once the null values have been eliminated. In a subsequent step, we narrow down the records to those with transcription lengths greater than 50. B. Initial Steps 1) Prevent the nltk library and domain-specific word removal: 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. Included in the Natural Language Toolkit (nltk) library is a collection of frequently used stop words that Tokenization with the help of the nltk library: The term "tokenization" describes the process of dividing a larger piece of text or a phrase into smaller pieces. A number of tokenization techniques are available in the nltk library. For example, there is word tokenize, which separates text into words, and message tokenize, which separates text into sentences. Sorting nouns using the nltk library for POS tagging: Basically, it's the act of breaking down a phrase into its component elements and giving each one a name. In order to filter for nouns and other particular parts of speech, the nltk library offers capabilities for doing POS tagging. When doing additional research on a text, this might help to isolate the most important terms. 4) WordNetLemmatizer for Lemmatization: Lemmatization is a technique for reducing words to their root form, which may increase the performance of natural language processing models and decrease the dimensionality of text data. To apply lemmatization to text data, one may make use of the WordNetLemmatizer that is part of the nltk package. 5) Eliminate duplicate features from the feature list: Over fitting and poor model performance might result from feature duplication. Hence, it's critical to eliminate characteristics that are duplicates. 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) TfidfVectorizer for factorization: ML models use factorization, which is the transformation of text input into numerical form. Using the term frequency inverse document frequency (TF-IDF) approach, which prioritizes less frequent terms in the text data, the TfidfVectorizer can victories text data.

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

That way, natural language processing models can function better.

Fig. 3. Word Cloud for Medical Transcripts

Preparing Data for the Model (C) 1) Split the data in half for training and testing purposes. This means we'll use half for practice and half for the real thing. 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 under Sampler for Under Sampling: Under sampling is a machine learning strategy that helps deal with class imbalance. This happens when there is a significantly higher prevalence of one class in the dataset compared to the others. Because of this, biased models may end up underperforming when applied to minority groups. You may use the RandomUnderSampler function from the imbalanced-learn Python package to systematically eliminate samples from the class with the largest number of members until the distribution of the classes becomes more balanced. If this is done, the model's performance on the minority class can be improved. Section 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. If you want to increase your model's performance, you may try using forward feature selection, which is adding one feature at a time from a blank slate. Training a model using each feature separately and then picking the best performing one is the first step. The procedure is then repeated by training models with every conceivable combination of the attributes that were previously chosen, and the best performing combination is picked. Iteratively, this procedure is carried out until the required performance level or the specified number of features is met. A number of benefits accrue from using forward feature selection, including its computational efficiency and the ease with which it reveals the dataset's most salient characteristics. Nevertheless, over fitting 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.

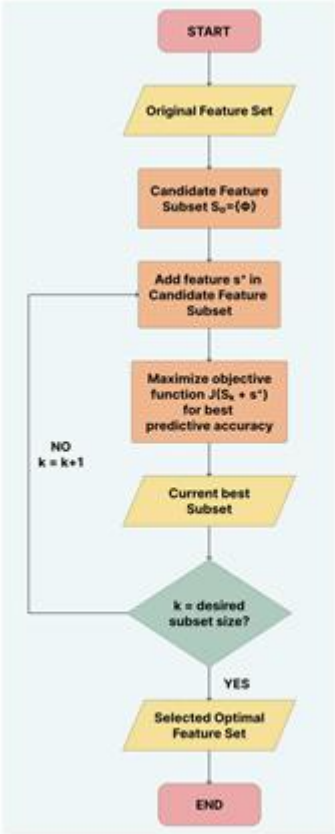

Fig. 4. Flowchart for Sequential Forward Selection (Feature Extraction Algorithm)

Part E: Creating Models 1) one statistical model that aims to estimate the likelihood of an outcome is logistic regression. To put it simply, it models the connection between a groups of independent variables. When dealing with difficulties involving several classes, logistic regression may be a useful tool. 2) Support Vector Machines (SVMs): These are supervised learning algorithms that are used for regression and classification purposes. Using a high-

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

dimensional space, support vector machines (SVMs) locate a hyper plane that effectively divides the data points into their respective classes. By converting the input data into higher-dimensional feature spaces, SVMs are able to manage non-linearly separable data. This is achieved via the use of non-linear kernel functions. 3) Random Forest: Used for feature selection, regression, and classification, it is a well-liked ensemble learning approach. 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 resilient to outliers and noisy data, and it can process a high number of input characteristics. Gradient boosting trees are a fourth kind of boosting technique; they take several poor models and strengthen them by training each one repeatedly to fix the mistakes made by the preceding one. Decision trees are used as the weak models in Gradient Boosting trees, an extension of Gradient Boosting. 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. Known for its capacity to manage intricate and non-linear connections between the input data and the target variable, Gradient Boosting trees are applicable to both regression and classification applications. 5) Gradient Boosting with a Categorical Focus: This approach was developed with categorical data in mind. A combination of categorical encoding methods like target encoding and decision trees with categorical splits, as opposed to continuous splits, is what makes categorical boosting work. One of categorical boosting's well-known uses is in classification problems; it excels at dealing with unbalanced datasets and features with a high cardiac index.

EXPERIMENTATION

During the data processing phase, we tried out several methods for data reduction in order to facilitate feature selection. One of these methods included using POS tagging to isolate nouns from transcripts. • 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 under fitting and 20 features was over fitting; 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 Regression, 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

The project's findings, including any new insights and their relevance to the initial business challenge, should be detailed in this section. • We said at the end of the research that there isn't enough data to make accurate forecasts and those additional records are needed. For the medical conditions "Surgery" and "Consultation and History," we were able to get a 99% accuracy rate using Categorical Boosting.

	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

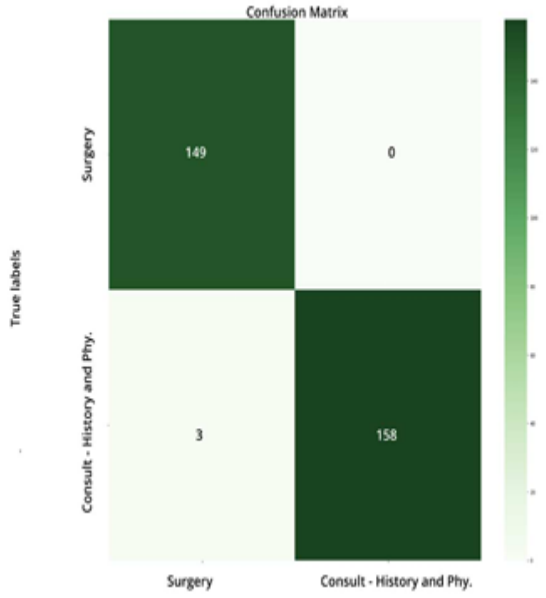

Fig. 6. Confusion Matrix (Heat map)

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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%. Because numerous transcripts from other courses describe surgeries related to medical specialties or patient histories, the data are being incorrectly mapped to the Surgery and Consultation and History classes. As an example, transcripts pertaining to cardiovascular and pulmonary patients may include information on heart surgeries or their medical history.

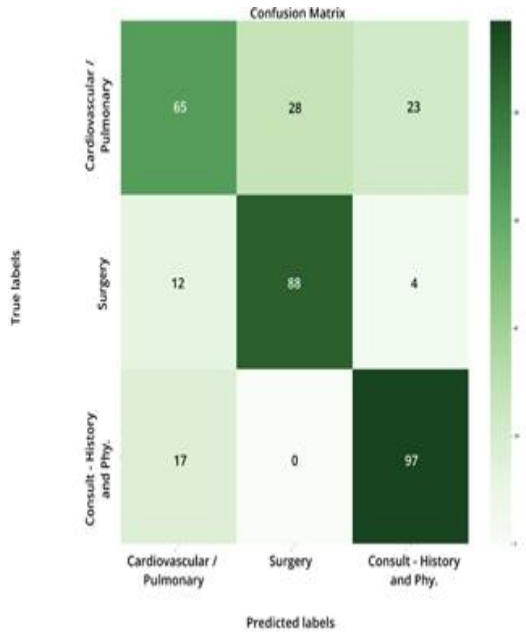

Fig. 7. Confusion Matrix (Heat map) for 3 Medical Illnesses using Categorical Boosting

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

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. Though manually creating features could improve the dataset's performance, these characteristics might 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. In order to tackle this, moving forward, we will segment the transcriptions into more precise pieces. So, we may classify the medical material with additional depth and granularity using a multiclass classification technique. We anticipate this will greatly enhance the precision of our categorization outcomes, leading to a more practical and dependable medical transcribing system. 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]. 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. Journal of biomedical informatics, 46(5), 869–875, (2013).

[5]. Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. What can natural language

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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

[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)

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).