# Group 5: Speech to text processing in Medical field

Madhvikaben Bhatt 1154135
Khyati Patel 1172720
Osheen Baby Varghese 1168517
Dhruvisha Patel 1159961
Niharika Sojitra 1170232

## Abstract

*In order to provide better and faster treatment to patients, NLP methods and AI models are greatly needed in the medical field. Since during pandemic speech recognition emerge exponentially, still there isn't much work done in the field of medical text classification, which may assist hospital professionals treat patients more efficiently, we came across the speech-to-text processing issue for this project. In this study, we attempted to give patient prescriptions in the form of speech, which we subsequently converted to text using NLP text processing techniques and models, which were then used to identify the disease category. By reducing unnecessary physical contact with documentation and form completion, this approach will save doctors' time when writing extensive medical prescriptions and lessen the risk of disease spread. Multinomial DB, Random Forest, SVM, and KNN had testing accuracy of 57.88%, 66.43%, 69.55%, and 65.72%, respectively.*

## 1 Introduction

In medical profession, text classification has significant impact on various application like obtaining knowledge of clinical results documented in the medical literature, diagnosis of diseases, medical research, and the automatic construction of disease ontology. Medical text classification is challenging due to the presence of terminologies that define medical concepts and terminologies. Furthermore, the medical data is frequently deficient in grammatical sentences and does not follow natural language grammar. Because the procedures for extracting text and training sets are different, the findings for text classification differ from those for medical text classifications.

During covid-19 pandemic, the need of automated speech recognition increased for various medical sectors to avoid physical contact. Since primary focus of prior research in this area is on textual data, we decided to take a step further, we introduce this idea of speech to text processing. We aim to recognise disease categories from human speech of medical prescription based on numerous NLP technologies. The figure given below depict our work process.

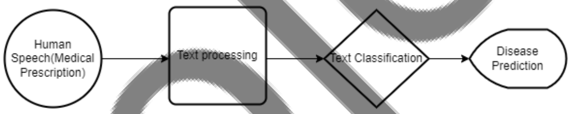


Figure 1: Input and output for the project

As shown in the figure 1, human speech, i.e., medical prescriptions, is provided as the input for our project. This voice is then transformed into text using existing APIs and libraries and further processed to identify illness symptoms. We are utilizing model predictions to forecast the disease class based on these extracted symptoms. We used the MultinomialNB, Random Forest, KNN, and SVM algorithms to train four models.
We conducted a manual analysis of 50 medical prescriptions with the help of a medical specialist to guarantee the model's accuracy. To test all of the models, we used this prescription as a set. SVM is the best model with the most correct sickness class predictions, while MultinomialNB predicts the least number of classes correctly.

## 2 Related work

Nafiz et al. feel that structural interventions can offer 92 percent accurate genuine and trustworthy data. The author's conclusions are skewed due to data collecting and analytic limitations. They have been unable to test further permutations due to data constraints, resulting in algorithm biases Sadman et al.(2020). To overcome constraints, by using manual coding or a word statistical approach, for the first time Xieling et. al. performed word-by-word examination method on a large bibliometric analysis of the present corpus of knowledge. However, there is still a need to improve and ensure the quality of life in clinical trials using NLP methodologies Chen et al.(2020).

Viincenza et al. worked on an application involving the digitization of medical prescriptions and the authorisation of health care services. However, they employed regular expression and string matching algorithms, which would not yield appropriate results since the medical text has distinct sentence structure and a lack of essential POS terms Carchiolo et al.(2019). Alexandre Trill used the similar method to extract the key phrases from the prescription for his project Trilla(2009). He employed regular expressions to extract symptoms and patient information from a medical prescription.

A document is viewed as an unordered collection of independent words having one or more occurrences, according to the BoW paradigm. The BoW assumption is employed in several widely used models, including tf-idf and Okapi BM25 based VSMs and Language Models Chowdhury (2010), Zhao and Mao (2017), Tsai (2012). A document is denoted as a numeric vector if a word is represented by a sequence of numeric numbers. The similarity determined by the similarities between query phrases and documents, as the query may be thought of as a combination of terms Jelinek(1980), Zhai and Lafferty(2017).

Cohan et al. proposed a model to identify damage events induced by medical errors in patient care and classify them according to their severity levels, which range from a dangerous condition to death, in another paper Sadman et al.(2020). He used an attention mechanism to increase the recurrent layer's performance in RNN model. However, due to the relative quantity of risk instances relative to the overall number of data in the category, the performance may have been improved even more if some categories were balanced in terms of harm and no-harm cases.

Gunjan Dhole et al. focuses on retrieving medical data from narrative clinical records using NLP Dhole and Uke(2019). This work demonstrates how NLP is used to extract medical information from a narrative text using tokenization, noun entity recognizers, parts of speech taggers, and connection extractors. The system proposed in this study tries to understand text related to a list of symptoms and then responds appropriately.

Although some work has been done in the medical field to classify text, it currently needs a speech recognition component for medical terminology. As a result, we chose to include it to supplement the earlier effort.

Table 1: Class labels and corresponding description of the disease

| Label | Description |
|---|---|
| 1 | Neoplasms |
| 2 | Digestive system |
| 3 | Nervous system diseases |
| 4 | Cardiovascular diseases |
| 5 | General pathological conditions |

## 3 Methodology

### 3.1 Data collection

On the internet, finding medical data in the correct format is tough. The bulk of the datasets we've come across have already been processed and have numerical data for the columns, which would be inconvenient for us because we needed textual data to test NLP techniques and models. However, on the Kaggle platform, we uncovered a dataset named "Medical Text" containing 14438 entries. Chaitnya(2018). Only two columns, label and medical prescription are present in the dataset. Every prescription in the range of one to five has a "label" connected with it. Table 1 shows our dataset related labels and descriptions.

### 3.2 Dataset balancing

It's crucial to balance dataset while developing a model. When a dataset is imbalanced, certain classes will have a large number of records while others will have a small number. Because the bulk of the class label will be utilised to forecast the label for each class, this will present a difficulty during model training. The random oversampling Brownlee(2015) approach was used because dataset used was uneven. This method produces multiple duplicate entries in order to make the minority and majority classes equal. As shown in the figure 2, records for class two are very less than class five. As shown in the figure 3, after random oversampling all the classes have 2477 records as the fifth class has, the same number of records.

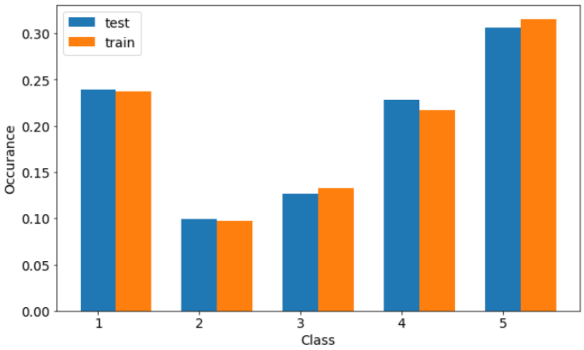
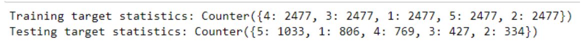


Figure 2: Class imbalance in the dataset

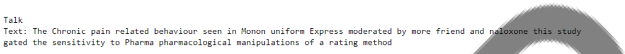


```
Training target statistics: Counter({4: 2477, 3: 2477, 1: 2477, 5: 2477, 2: 2477})
Testing target statistics: Counter({5: 1033, 1: 806, 4: 769, 3: 427, 2: 334})
```

Figure 3: Balanced training dataset

### 3.3 Speech to text conversion

To accomplish this, there was use of pipwin, pyaudio, and SpeechRecognition APIs. These APIs identify and create text based on the speech delivered to the device. The example output from the speechrecognizer is shown in the 4. Even if the recognizer misses a few phrases, it successfully recognises the vast majority of medical terms.



```
Talk
Text: The Chronic pain related behaviour seen in Monon uniform Express moderated by more friend and naloxone this study
gated the sensitivity to Pharma pharmacological manipulations of a rating method
```

Figure 4: Speech to text conversion

### 3.4 Text Pre-processing

It is self-evident that when working with text data, the data must be cleaned and preprocessed before being used for further research. The nltk library was used to complete this step. Stopwords were eliminated, the content was rewritten in lowercase, and hyperlinks, punctuation, and numerals were removed. We used the SpellChecker library to repair the incorrect words. Then text was tokanize for further processing. After tokenizing the text, lemmatization was done which takes the context into consideration when converting a word to its meaningful base form, known as a lemma.

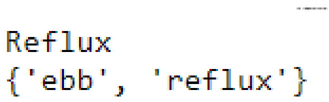
#### 3.4.1 Vectorization of the text

Textual content, which we must work with, is incomprehensible to machines. Due to this , translating input textual information into vectors is a necessity. As a result , there is need to build vectors of phrases in the texts, together with their associated frequencies. A document represents each medical prescription in the collection. There was use of a count vectorizer, a bag of words(BOW), and TF-IDF vectorization.

BOW and count vectorizer is just representation of unique words in term of frequency, while the TF–IDF value is given by number of times a word appears in a document and is offset by the number of documents in the corpus that contain the term, which helps to account for the fact that some words appear more frequently than others in general. The term-weighting system, TF–IDF is one of the most prevalent technique today.

#### 3.4.2 Word similarity check

Because certain medical terminology are not often used, it can be difficult to determine their meaning or create a disease link. In previous study, it was determined that phrases with the same amount of characters are considered synonyms. This isn't always the case, though. To tackle this problem, we used the Wordnet database to find synonyms for the terms. Because nouns make up the bulk of symptoms, we extracted nouns from the text and used them as an argument in the wordnet's synsets function to find synonyms. All feasible combinations were made, using these synonyms to identify the prescription's class label. As we can see in the above image 5,all the possible synonyms of certain term received using the wordnet library.



```
Reflux
{'ebb', 'reflux'}
```

Figure 5: Synonyms for the word

### 3.5 Model preparation

For modelling, we used the sklearn package. This library includes classes for all of the built-in algorithms. We used the MultinomialNB, Random Forest, SVM, and KNN algorithms to train our models. We utilised a 70:30 split for the workout and test. To exploit the built-in preprocessing capabilities, we used the ski-kit learn library, which offers the Pipeline class, which handles all of the work for us.

#### 3.5.1 MultinomialNB

For classification using discrete features, the multinomial Naive Bayes classifier is appropriate Shriram(2021). It uses the strong Naive Bayes assumption that each feature is independent of the others to predict a sample's category. As a consequence, we can compute the probability of each category, and the category with the highest probability will be the output.

$P(A/B) = P(A) * P(B/A)/P(B)$

This method is simple to design, a good choice for text classification with quick processing, which is the main reason we chose this technique for training a model.

However, there are a number of drawbacks to using this approach, including its lower accuracy when compared to other algorithms and inability to handle numeric data. We used four different strategies to look into the disparities in model predictions.

#### 3.5.2 Random Forest

Random forest Wikipedia(2014) is an ensemble learning method that uses a number of decision trees to solve classification, regression, and other problems. For classification tasks, the random forest's output is the class picked by the most trees. Because it aggregates all of the outputs into a single result, it has a high level of accuracy. Over-fitting is to be reduced, resulting in minimal variance and good accuracy. It does, however, have a lot of trees, which increases its computational complexity.

#### 3.5.3 Support-vector machine

Support-vector machinesWikipedia(2014) are the supervised learning models using learning algorithms that evaluate data for classification and regression analysis in machine learning. This algorithm offers good accuracy and performs faster prediction than the Naive Bayes. When the number of features exceeds the number of samples, this technique is memory economical and adaptable, but it also might lead to over-fitting.

#### 3.5.4 K-Nearest Neighbour

Class membership is the result of k-NN categorization Wikipedia(2014). An object is categorized based on a majority vote of its neighbours, with the object being allocated to the most common class among its k closest neighbours (k is a positive integer, typically small). If k = 1, the object is simply assigned to the nearest neighbour's class. We have selected this algorithm with K=5, as it achieves high accuracy in a wide variety of prediction type problems.

Table 2: Quantitative result

| Model | Training accuracy | Testing accuracy |
|---|---|---|
| Multinomial NaiveBayes | 64.15 | 57.88 |
| Random Forest | 100 | 66.43 |
| SVM | 82.95 | 69.55 |
| KNN | 68.43 | 65.72 |

## 4 Results

We examined the model accuracy after and after text preprocessing and to see if text preprocessing has an impact on the result. Text preprocessing results in 3% rise, according to our findings.

### 4.1 Quantitative Results

According to the table 2, SVM had the highest accuracy in the testing set, with 69.55 percent. Random Forest is in second place, with a testing accuracy of 66.43 percent and a training set accuracy of 100 percent. With 68.43 percent and 57.88 percent, respectively, the KNN and MultinomialNB models came in last.

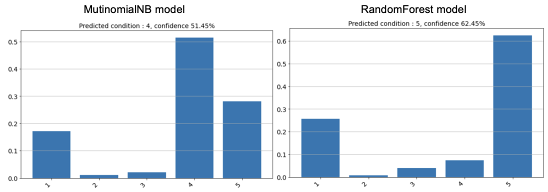


Figure 6: prediction for the prescription : i)multinomialNB Model ii)RandomForest model

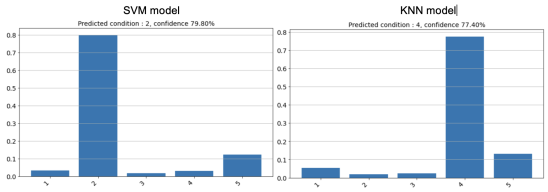


Figure 7: prediction for the prescription : i)SVM Model ii)KNN model

The term "confidence" refers to the model's belief that the provided prescription will be labelled with that precise label. As manifested in the Figure 6, the multinomialNB model correctly predicted the class of the provided medical prescription with confidence of 51.45% which was least while SVM

Table 3: Qualitative analysis result

| Model | Result |
|---|---|
| Multinomial NaiveBayes | 26 |
| Random Forest | 29 |
| SVM | 31 |
| KNN | 28 |

predicted correct labels with highest confidence of 79.80% which is shown in Figure 7.

### 4.2 Qualitative Results

We have performed a manual analysis of 50 randomly selected prescriptions to evaluate the model results. We have allocated labels to each of these prescriptions with the help of a knowledgeable medical field person and also the internet. After training our models, we applied the same 50 prescriptions set to each model and outcomes are expressed in Table 3. According to Table 3SVM predicts the prescriptions with the highest accuracy. Random Forest, KNN, and MultinomialNB have accuracy in respective orders.
This also proves that our quantitative findings were precise.

## 5 Limitations and Future work

### 5.1 Limitations

There are some limitations to our work that needs to be addressed. First, the speech recognizer we used stops after 25-30 words of continuous word listening and generates false text for the given audio. Second, noise removal has yet to be completed, which could explain why the speech recognizer failed to reliably distinguish certain voiced words. Furthermore, our algorithms have only been taught to predict five diseases, a number that should be expanded as new disease categories are added. In order to enhance their accuracy, models must be trained with more records, as a single inaccurate prediction in this domain could result in a considerable loss of life.

### 5.2 Future work

For the future paths, we've opted to focus on enhancing voice recognition to correctly forecast speech and eliminate noise. The medical prescription will be input into the application platform (GUI) and classified. We also wish to focus on the challenge of removing noise from the input speech. Furthermore, we wish to use the LSTM, RNN, and BERT algorithms to train models because they are often used in the articles we have studied for the project literature. In addition, we wish to employ a genetic algorithm to create accurate and efficient forecasts.

## 6 Conclusion

If additional research is done, AI and NLP might play a crucial role in the medical field. Using the SVM model, which is the best of the four models, we got about 70% accuracy for voice to text conversion and 69.55 percent accuracy for prediction in our experiment. While Random Forest achieved 100% training accuracy, it only achieved 66.43 percent testing accuracy. We used a range of tools and libraries to fulfil our tasks, such as tokenization and language models, which will be essential in future, as NLP has a promising future in AI and model building for medical industry.

## References

Jason Brownlee. 2015. Random oversampling and undersampling for imbalanced classification.

Viincenza Carchiolo, Alessandro Longheu, Giuseppa Reitano, and Luca Zagarella. 2019. Medical prescription classification: a nlp-based approach. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 605–609. IEEE.

Chaitnya. 2018. Medical text.

Xieling Chen, Haoran Xie, Gary Cheng, Leonard KM Poon, Mingming Leng, and Fu Lee Wang. 2020. Trends and features of the applications of natural language processing techniques for clinical trials text analysis. *Applied Sciences*, 10(6):2157.

Gobinda G Chowdhury. 2010. *Introduction to modern information retrieval*. Facet publishing.

Gunjan Dhole and Nilesh Uke. 2019. Medical information extraction using natural language interpretation.

Frederick Jelinek. 1980. Interpolated estimation of markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980.*

Nafiz Sadman, Sumaiya Tasneem, Ariful Haque, Md Maminur Islam, Md Manjurul Ahsan, and Kishor Datta Gupta. 2020. "can nlp techniques be utilized as a reliable tool for medical science?"-building a nlp framework to classify medical reports. In *2020*

*11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0159–0166. IEEE.

Shriram. 2021. Multinomial naive bayes explained.

Alexandre Trilla. 2009. Natural language processing techniques in text-to-speech synthesis and automatic speech recognition. *Departament de Tecnologies Media*, pages 1–5.

Chih-Fong Tsai. 2012. Bag-of-words representation in image annotation: A review. *International Scholarly Research Notices*, 2012.

Wikipedia. 2014. Random forest.

Chengxiang Zhai and John Lafferty. 2017. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM New York, NY, USA.

Rui Zhao and Kezhi Mao. 2017. Fuzzy bag-of-words model for document representation. *IEEE transactions on fuzzy systems*, 26(2):794–804.

## 7 Appendix

### 7.1 Softwares and libraries used

Anaconda (Jupyter notebook): Platform to implement project in python
SpeechRecognition: To recognize given speech
pyaudio: Library to support SpeechRecognition
pandas: Library to deal with dataframe objects
nltk: Library to deal with text processing, and text operations like POS tagging, vectorizing
matplotlib: Library to visualize the results
spellchecke: Library to correct the spellings of the words
sklearn: Library to use inbuilt model functions
imblearn: To balance the dataset
pipwin: Supportive library for the SpeechRecognition

### 7.2 Group Member Contribution

**Madhvikaben Bhatt**
She has done duties for literature review, manual analysis, text pre-processing, algorithm implementation and training models for classification and project related report writings.

**Osheen Baby Varghese**
She has worked on Literature review, data pre-processing and text vectorization.

**Khyati Patel**
She has done Literature review, Identifying limitations of the previous work and qualitative analysis.

**Dhruvisha Patel**
She worked on Speech to text feature, Identifying limitations of previous work.

**Niharika Sojitra**
She has done Speech to text feature,Identifying limitations of previous work and report writing.