Symptom Extraction and Disease Classification from Patient's Description

Xiaoliang (Toby) Zhu [1]

Abstract

Background Chronic obstructive pulmonary disease (COPD) is a lung disease with a high lethal rate whose symptoms are similar to typical pulmonary diseases. Many COPD patients can hardly notice those symptoms or mistake them for symptoms of other diseases, thus receiving inadequate medical treatment in its early stage.

Objective We hope to distinguish COPD symptoms from signs of other non-COPD pulmonary diseases by analyzing patients' textual data while simultaneously extracting corresponding symptoms. We want to develop a chatbot to communicate with patients and help them identify COPD-related symptoms.

Methods We build a machine learning pipeline consisting of a binary classification model and a Named Entity Recognition model. We try models like KNN, bi-LSTM, and BERT to conduct the binary classification and employ pre-trained BERT models to deal with the NER task.

Results We found that the BERT model performs better than the bi-LSTM model for classifying symptoms, while the two pre-trained models perform about the same in the NER task. The chatbot is deployed using python Flask with the COPD classification function.

Discussions & Conclusions Our research goal is achieved with some limitations. The chatbot pipeline is constructed without the NER function due to library incompatibility. Although the size of the entire dataset is small with significant unbalances, this project demonstrates the potential of applying the deep learning pipeline to help diagnose symptoms of early-stage diseases. We will finish the deployment in future works and extend our model to other categories of diseases for classification and symptom extraction.

[1]Vanderbilt University, Nashville, Tennessee, USA. Correspondence to: Xiaoliang Zhu <xiaoliang.zhu@vanderbilt.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

1. Introduction

In this section, we deliver the general background knowledge for this paper, stating the research problem and conducting a brief survey on the existing literature. In the end, we have a summary of our methods and key findings. Further implications of our works would be covered as well.

Background Chronic obstructive pulmonary disease (COPD) is a chronic inflammatory lung disease that blocks the airflow from the lungs. Its symptoms include cough, breathing difficulty, mucus production, and wheezing. The typical cause of COPD is long-term exposure to irritating gases or matters, mostly from cigarette smoke.

COPD is infamous for its vulnerability and incurability. First of all, COPD is not curable. Its symptoms and progression can only be delayed. It can progressively worsen with everyday activities such as walking or dressing becoming difficult. Research showed that around 4% of COPD patients died during hospitalization, and about 22% of hospital survivors died one year after discharge (Ho et al., 2014). Moreover, people with COPD have relatively higher risks of developing heart disease, lung cancer, and many other conditions. In 2019, COPD caused around 3.2 million deaths (WHO, 2021).

Challenges and Research Problem A significant lethal factor of COPD is that its main symptoms are similar to those of usual pneumonic diseases, such as bronchitis or pneumonia. Hence, many patients may not even notice the severity of their symptoms. Some statistics showed that around 80% of patients have already entered stage III, the late-stage COPD, before their diagnosis (Lee et al., 2016).

Due to its hiddenness, the early diagnosis of COPD became an important topic. Typical approaches involve the use of spirometry and COPD assessment tests. However, spirometry can lead to an over-diagnosis of COPD in the elder population (Qaseem et al., 2011), and both measures

are usually insufficient for the accurate diagnosis of COPD. Other tests like chest X-rays and CT scans are useful only in excluding other conditions or including some comorbidities, but not for confirming the diagnosis.

Recent advances in healthcare have made many uses of the Natural Language Process (NLP) techniques. People employed text mining methods on patients' electronic health records (EHR), trying to identify pertinent and potentially complex information about symptoms and phenotypes from the textual data. Regarding the similarity between symptoms of COPD and other pulmonary diseases, it would be natural for us to ask if we may distinguish COPD from other pulmonary diseases through textual data from patients and, if possible, identify corresponding symptoms simultaneously. By solving this problem, we would obtain a powerful approach to diagnosing COPD in its early stage, which may save thousands of lives in the world.

Existing literature Many existing works of literature focus on predicting the mortality in COPD using different indices such as functional dyspnea, BMI, and FEV_1 (forced expiratory volume in one second) as predictors. For example, Marin et al. (2013) showed that ADO (age, dyspnoea, and FEV_1), BODE (BMI, airflow obstruction, dyspnoea, and exercise capacity), and e-BODE (BODE plus exacerbations) are the best indices to predict 6-month mortality of COPD. Other studies tried to predict COPD from factors like symptoms and blood test indices. For instance, Ohar et al. (2010) assessed the airflow obstruction and respiratory symptoms in people with a smoking history, concluding that COPD-related symptoms are frequent in people with airway obstruction, which would increase their risks for COPD. Xiong et al. (2017) also demonstrated that parameters such as neutrophil-to-lymphocyte ratio (NLR) and eosinophil-to-basophil ratio (EBR) in routine blood tests (RBT) correlate significantly with the severity of patients with COPD.

While studies mentioned above utilized clinical testing data, some researchers used EHR to predict COPD. Himes et al. (2009) predicted COPD on patients' EHR with an accuracy of 0.833 using a Bayesian network model composed of age, sex, race, smoking history, and eight comorbidity variables. Moreover, Agarwal et al. (2017) employed the Clinical Text Analysis and Extraction System (cTAKES) to extract features from clinical notes and make annotations, conducting COPD classification tasks with traditional machine learning methods like Naïve Bayes, Random Forest, and Support Vector Machine. The best classifier was Random Forest, with an AUC of around 0.693.

While previous research achieved fruitful results, few researchers have experimented with predicting COPD with the deep learning approach. As the artificial neural networks (ANNs) become increasingly powerful in recognizing patterns and making predictions, modern networks like Recurrent Neural Networks (RNNs) or Transformers may offer a more accurate prediction of COPD. Therefore, we incorporate many deep learning models in our project in order to have a more thorough analysis of textual data from patients.

Summary In this work, we predict if a patient has gotten COPD by analyzing the patient's description of symptoms. Beyond that, we try to extract the specific symptoms of the diseases embedded in the provided description. We experiment with different models, including k-Nearest Neighbors (kNN), bi-directional LSTM (bi-LSTM) model, BERT model, and other classic NLP models. The data consist of textual data from research papers and patient-doctor conversations from online health Q&A platforms. We find that the BERT model performs significantly better than the bi-LSTM model, while the two pre-trained BERT models that we employed have a similar performance. Finally, we create a chatbot that can determine if a patient has gotten COPD according to the provided description of this patient's symptoms. Our work demonstrates the potential of using the deep learning neural network structure for early-stage COPD diagnosis, and our chatbot is a prototype of automated symptom self-identification for COPD. It can help decrease the cost of disease diagnosis and make COPD detection more convenient. Overall, it may help patients diagnose early-stage COPD, thus boosting their possibility of survival.

2. Pipeline and Data

In this section, we discuss the data collection and detail the research problem. The research pipeline would be mentioned as well.

2.1. Research Problem

We first state our research problem in detail. The problem that we are interested in is how we can distinguish symptoms of COPD from other non-COPD pulmonary diseases, and our goal is to construct a machine learning pipeline that detects whether the patient has gotten COPD by analyzing textual descriptions. Furthermore, we hope this pipeline can simultaneously extract corresponding symptoms from the descriptions so that the pipeline can provide a quick self-check for the patient and help detect the early-stage COPD symptoms. Thus, we break this problem into two parts, one modeled as a binary classification problem and the other as a Named Entity Recognition (NER) problem. We try both supervised and unsupervised approaches for the binary classification task, and we use the supervised learning method only for the NER task.

2.2. Research Pipeline

Figure 1 displays our research pipeline. The input of this pipeline is the patient's description of the recent symptoms consisting of several English sentences. Those sentences will be fed into our first model - the binary classification model - after being split and cleaned. The model would then compute the probability that this patient has gotten COPD according to phrases and structures embedded in given sentences. After outputting the probability, those sentences would then be sent to the NER model so that it can truncate the sentences and extract phrases that describe the patient's symptoms. In the end, our pipeline would generate a probability indicating whether this patient has gotten COPD and attach symptom phrases extracted from the input text.

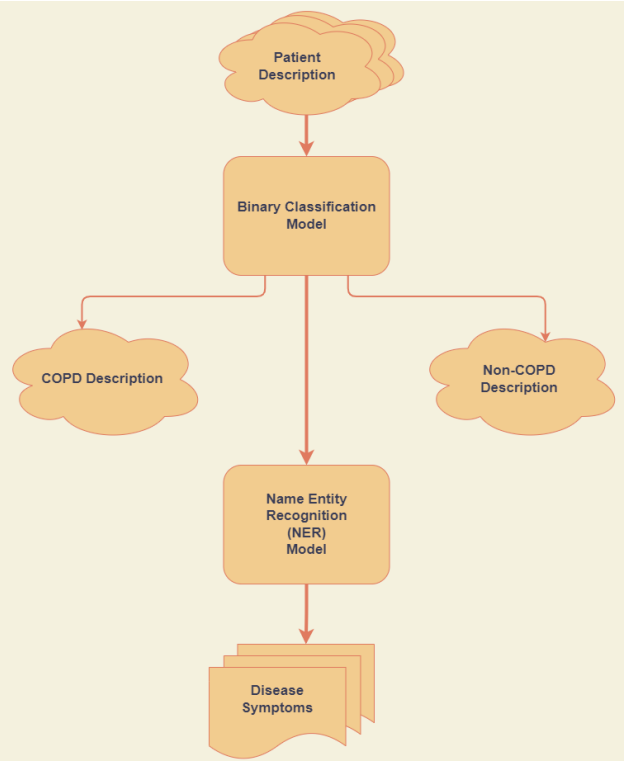


Figure 1. Research Pipeline. The pipeline would first classify the input text into COPD or Non-COPD category, and then extract symptom-related phrases from those sentences.

2.3. Data

We have two datasets to train the binary classification model, the COPD corpus and the COVID-Dialogue-Dataset. The COPD corpus (Ju et al., 2019) is a semantically annotated corpus focusing on the phenotypic information of COPD and consisting of 30 full-text articles. Its training portion contains 6219 sentences with an average length of 11 words (Fig. 2). The COVID-Dialogue-Dataset (Zeng et al., 2020) is a bilingual medical dialogue dataset about COVID-19 and other types of pneumonia. It contains conversations between doctors and patients concerned that they may be infected by COVID-19 or other pneumonia. We select the English dataset that contains 1401 informative consulting sentences from three online Q&A-based medical advisory platforms, and the average length of each sentence is about 24 words (Fig. 2). The sources for both datasets are detailed in the reference.

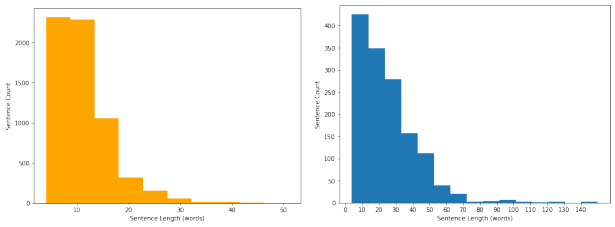


Figure 2. Left: distribution of sentence length in the COPD corpus. Right: distribution of sentence length in the COVID-Dialogue-Dataset. Although the COPD corpus contains more sentences, its average sentence length is smaller than that of the COVID-Dialogue-Dataset.

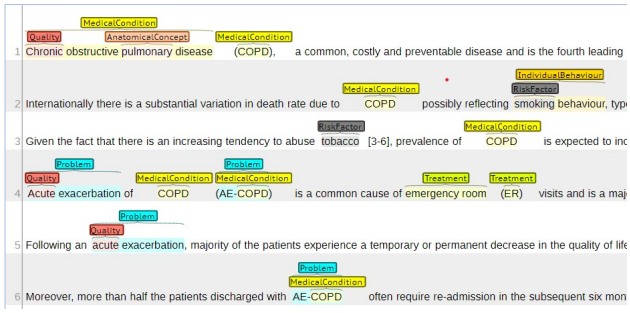
For the NER model, we use the National Center for Biotechnology Information (NCBI) Disease Corpus (Doğan et al., 2014), a collection of fully annotated paper abstracts at both mention and concept levels. It is a classic dataset for training a disease-recognition model, containing 6892 disease mentions and 790 unique disease concepts. We choose this new dataset to train the NER model instead of the previous two datasets because we want our model to have the ability to extract symptoms of different types of disease categories, instead of focusing on respiratory diseases only. Moreover, the COVID-Dialogue-Dataset does not have corresponding annotations. If we train our NER model solely on the COPD corpus, the model can have awful performance in extracting symptoms of other pulmonary diseases. The source of the NCBI Disease Corpus is included in the reference as well. Figure 3 is an illustration of those three datasets.

3. Methods

In this section, we discuss the methods and models used to solve this research problem. We first introduce our approach to deal with the binary classification task and then the NER task. We will consider the choice of statistical inference model in the next section.

3.1. Binary Classification for COPD & Non-COPD

We experimented with both the unsupervised learning and the supervised learning approach for the binary classification task. Before utilizing deep learning models, we also try a



id=1
https://www.icliniq.com/qa/covid-19/i-have-cough-with-no-travel-history-is-this-a-symptoms-of-covid-19

Description
I have cough with no travel history. Is this a symptom of Covid-19?

Dialogue
Patient:
Hello doctor, I get a cough for the last few days, which is heavy during night times. No raise in temperature
Doctor:
Hello, I understand your concern. I just have a few more questions.
Does your cough has phlegm? Any other symptoms like difficulty breathing? Any other medical condition su
Patient:
Thank you doctor,
I have phlegm but not a lot. A tiny amount comes out most of the time. I have no difficulty in breathing. No
Doctor:
Hi, I would recommend you take n-acetylcysteine 200 mg powder dissolved in water three times a day. You

10090880|t|Mutation and haplotype studies of familial Mediterranean fever reveal new ancestral relationships
10090880|a|Familial Mediterranean fever (FMF) is a recessive disorder characterized by episodes of fever with
was 21%, with E148Q the most common mutation. Several lines of evidence indicate reduced penetrance amo

10090880	34	62	familial Mediterranean fever	SpecificDisease	D010505
10090880	200	228	Familial Mediterranean fever	SpecificDisease	D010505
10090880	230	233	FMF	SpecificDisease	D010505
10090880	240	258	recessive disorder	DiseaseClass	D030342
10090880	299	308	serositis	DiseaseClass	D012700
10090880	312	321	synovitis	DiseaseClass	D013585
10090880	327	330	FMF	Modifier	D010505
10090880	1087	1090	FMF	SpecificDisease	D010505
10090880	1522	1525	FMF	SpecificDisease	D010505

Figure 3. Upper: the COPD corpus. Mid: the COVID-Dialogue-Dataset. Lower: the NCBI Disease Corpus. The conversation records in the COVID-Dialogue-Dataset are from [icliniq.com], [healthcaremagic.com], and [healthtap.com].

traditional machine model to compare the performance. We would start with unsupervised learning.

Unsupervised Learning

When classifying a text between two categories, a natural idea would be to construct word vectors for each token and see if there forms any cluster. In this case, I first construct word vectors for the COPD corpus and the COVID-Dialogue-Dataset together using Word2Vec. Then, I use the KMeans clustering algorithm to cluster those word vectors, hoping that they can form separated word clusters that represent COPD-related phrases and non-COPD-related phrases. Since the task is a binary classification, we set the number of centroids to be 2. For visualization, I conduct a dimension reduction after clustering, compressing the word vectors into 2D dimensions. However, the clustering result looks nasty on the graph, as most data points mix together and are hard to separate. Applying a cosine kernel to represent the similarity among word vectors cannot separate them as well (Fig. 4). Although the Silhouette score for this clustering is not trivial (greater than 0.6), the score for one of the clusters is always negative. This result seems not reasonable, and our model does not adopt this model.

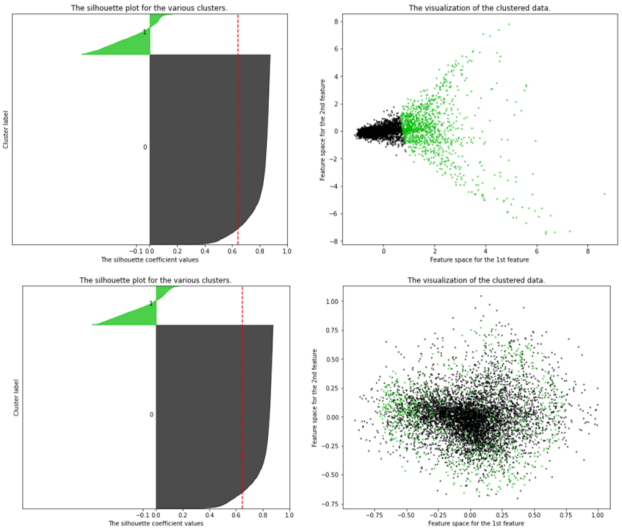


Figure 4. Upper: visualization of KMeans clustering with Silhouette analysis using PCA. Lower: visualization of KMeans clustering with Silhouette analysis using cosine-kernel PCA.

Supervised Learning

We propose the k-Nearest Neighbor (KNN) model, bi-directional LSTM (bi-LSTM) model, and BERT model for the supervised learning approach.

KNN Initially, we employ a traditional machine learning model, k Nearest Neighbor. We first construct word vectors for the COPD corpus and the COVID-Dialogue-Dataset respectively. Tokenizing every new input sentence into small tokens, we calculate the distance from each token to its k nearest tokens in the two datasets, respectively. In this case, for every sentence, we can sum those distances and get two final values, each representing the distance between this new sentence and one of the two datasets. We then classify this sentence into the category where it has a closer distance. We conduct this classification for each sentence from the patient's input and classify the entire text by the majority. Hence, the input feature is tokenized sentences, and the output is the classification result. The reason for choosing this model is intuitive, as it is natural to classify a text by checking the similarity of its sentences with sentences in other classified texts. If the text is similar to the COPD corpus, it is then plausible to be considered a COPD description.

For training and testing, we generate ten different splits for both datasets and experiment with this KNN model using different values of k from 5 to 30. For each of the ten splits, we first construct word vectors on the training split and test

the model on the test split. However, the result is not pleasant. The model reaches the best performance when k is 5, which only yields an average accuracy of around 18.62%. We calculate the accuracy as the proportion of correctly classified sentences among all testing sentences, which shows a poor model performance on prediction. Hence, the KNN model is not applicable in our pipeline.

Bi-directional LSTM Here we start utilizing the deep learning approach. As the text data is sequential and we hope to capture the temporal dynamics and context relationship within the data, we will use the RNN-based models. To avoid the exploding and vanishing gradient problem, we will adopt the Long Short Term Memory (LSTM) Networks with a bi-directional design so that the model can proceed in both forward and reverse orders while capturing both right and left context. Our bi-LSTM model has an embedding layer, a bi-directional LSTM layer, and a linear layer with the sigmoid function as the activation function. In this case, the input features are clean sentences from patients' descriptions, and the output vector, with a dimension of 2, indicates the probability that this sentence belongs to the COPD or Non-COPD description.

As noticed in the data section, a significant unbalance exists between the COPD corpus and the COVID-Dialogue-Dataset dataset, as the ratio of the number of sentences in the two datasets is about 6:1. To address this unbalance, we adopt the Cross-Entropy Loss as the loss function and assign a class weight to each class. The weight for each class is calculated by subtracting the proportion of each class in the entire dataset from one (1). This technique alleviates the unbalance and boosts the model performance significantly.

$$w_i = 1 - \frac{size\ of\ class_i}{size\ of\ entire\ dataset} \tag{1}$$

For the size of the embedding layer, we try different values among 200, 250, 300, and 350. The size of 300 gives the best performance. Next, we conduct 10-fold cross-validation to tune the hidden layer size (Fig. 5). We finally pick 35 as the layer size, as it exhibits the highest average F1 score and the lowest variance. We calculate the F1 score by taking the harmonic mean of the precision and recall. The reason for adopting the F1 score as the metric is that we want to take the unbalance of the data into account while including both precision and recall for an accurate evaluation. We finally generate ten different splits of our data where we train and test this bi-LSTM model, which yields an average F1 score of around 0.92.

BERT Classifier The bi-LSTM model works decently. However, as medical terminology can have different meanings according to the context, we hope the model can better utilize the surrounding text to establish context and capture the meaning of ambiguous language more precisely. Therefore, here we introduce the BERT model. In our model structure, after receiving the tokenized vectors from BERT, we pass the embedding vectors of the [CLS] token to a linear layer with the sigmoid function as the activation function. Similar to the bi-LSTM model, the input features are clean sentences from patients' descriptions, and the output vector, with a dimension of 2, indicates the probability that this sentence belongs to the COPD or Non-COPD description.

The method employed to alleviate the class unbalances is the same as the one in the bi-LSTM model. However, this raw class weight does not yield a great result. Hence, starting from this original class weight, we conduct 9-fold cross-validation to tune this assigned class weight (Fig. 5). We finally select the class weight of the COPD class to be 0.1439 and 0.8561 for the Non-COPD class, as it shows the highest average F1 score and the lowest variance. The F1 score is used as the metric for the same reason indicated before. We finally generate ten different splits of our data where we train and test this BERT classifier, which yields an average F1 score of around 0.95.

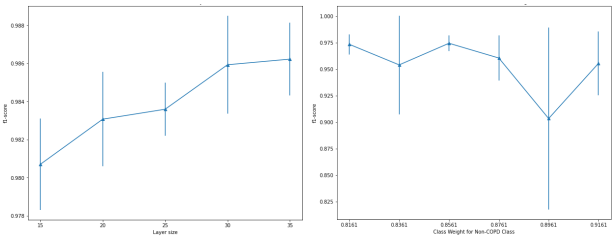


Figure 5. Left: F1 scores from cross-validation on hidden layer size of the bi-LSTM model. Right: F1 scores from cross-validation on the class weight of the BERT model. Each data point represents the average F1 score from a different split of data, and the vertical bar represents the variance.

3.2. NER for Symptom Extraction

For the NER task, we want to recognize and extract the disease and symptom information from the text. In this case, we also want a model that can understand the context embedded in the text and capture the deep relationship among tokens. Thus, we apply the BERT model again. We use the BertForTokenClassification class from the transformer package for token-level predictions, which wraps the Bert model with a linear layer to form a classifier. We tried two pre-trained models, the bert-base-cased and the biobert-base-cased, the latter of which is pre-trained on some medical-related NER datasets. The input features are tokenized texts with corresponding labels for each token. The output is the predicted label for each token in the input sequences.

We again generate ten different splits on the NCBI dataset and conduct experiments using the two BERT models. The

metric is accuracy, as the class unbalances here are not as severe as that in the dataset used for binary classification. The accuracy for both models is around 0.96, a remarkable model performance.

4. Result

4.1. Statistical Inference

We generate ten different splits from the corresponding datasets for each task and run experiments with different models. We then want to check whether the mean of metric values obtained is different or the same for each model, thus comparing the performance among distinct models. As we only have sample data, we will employ the paired-samples t-test, or dependent t-test, as the statistical model to test our hypothesis.

4.2. Binary Classification

We generate ten different splits from the COPD corpus and the COVID-Dialogue-Dataset and run experiments using both the bi-LSTM and BERT models on each split of the data. The bi-LSTM model produces an average F1 score of around 0.92, whereas the BERT model yields an average F1 score of 0.95 (Fig. 6). The null hypothesis is that the mean of model performances for the bi-LSTM model and BERT model are the same, while the alternate hypothesis is that the BERT model performs better than the bi-LSTM model. The resulting t-score is about 2.32, and the p-value is about 0.0455, larger than 0.001 but smaller than 0.05. Hence, although not very significant, we can claim that the BERT model performs better than the bi-LSTM model for this binary classification task.

4.3. Named Entity Recognition

We generate ten different splits from the NCBI Disease Corpus and run experiments using BERT with bert-based-cased and biobert-base-cased pre-trained models on each split of the data. Both give an average accuracy of around 0.96 (Fig. 6). The null hypothesis is that the mean of model performances for the two pre-trained models are the same, while the alternate hypothesis is that the performances are different. The resulting t-score is about 0.76, and the p-value is about 0.4674, much larger than 0.001. Hence, we fail to claim that the two models have any performance difference in this NER symptom extraction task.

4.4. Chatbot Deployment

We deploy the pipeline using python flask and build a static front-end webpage using HTML5 and CSS. As the PySpark and the Spark NLP libraries work abnormally on our local machine, currently only the binary classification model is deployed successfully on our chatbot. The chatbot's interface imitates the appearance of the chatting interface of WeChat. When the patient enters the interface of the chatbot, the bot will send a welcome message and introduce its function (Fig. 7). The limitation of 25 words is the truncating size of the model's word vectors.

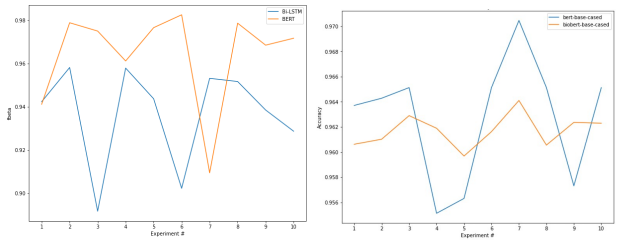


Figure 6. Left: the F1 score of bi-LSTM and BERT model on ten different splits of data. Right: the accuracy of BERT with two different pre-trained models on ten different splits of data.

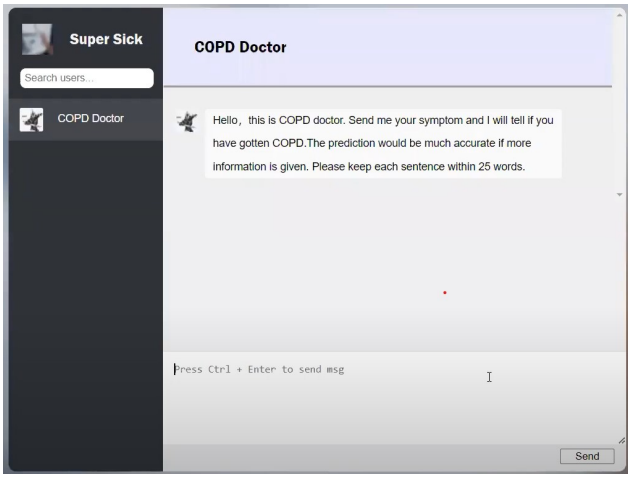


Figure 7. The chatbot's interface. Enter text into the message box below and click "send" to let the bot do the diagnosis.

If the patient enters nothing, the bot would reply with a warning. If the text entered is meaningless or has nothing to do with the disease, the bot would request more information. In the current version, if the patient enters any disease-related texts, the bot would convert them into tokens and feed those tokens into our pipeline. After a while, it will reply with the probability that the patient has gotten COPD (Fig. 8). In the future version, we plan to add the NER task after the classification task to extract symptoms from the input text so that the entire research pipeline is implemented and deployed.

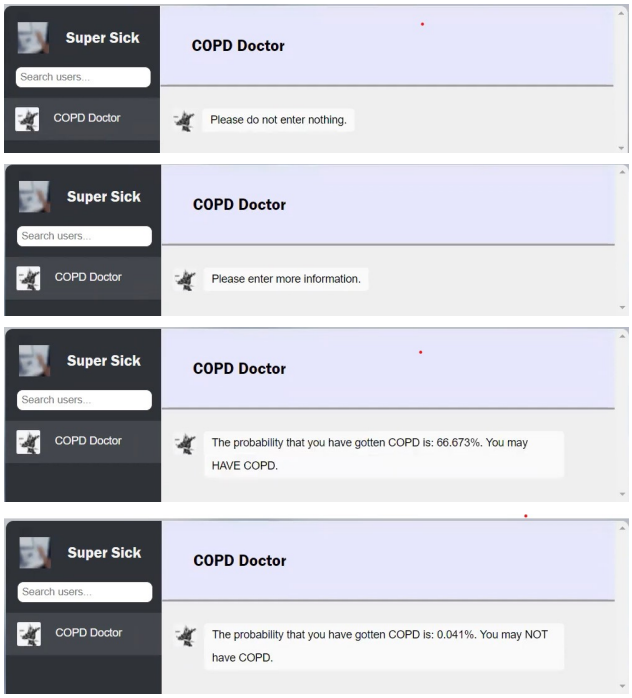


Figure 8. First: response if enter nothing. Second: response if too little information is provided. Third: diagnosis when input texts are classified as COPD. Fourth: diagnosis when input texts are classified as Non-COPD.

5. Discussion

The principal achievement in our work is the creation of that COPD chatbot. Moreover, we found that the BERT model performs better than the bi-LSTM model for classifying COPD and Non-COPD symptoms, while the two pre-trained models, bert-base-cased and biobert-base-cased, perform about the same in the NER symptom extraction task.

Our research goal is to distinguish symptoms of COPD and other pulmonary diseases by analyzing the textual data. We develop several models to conduct the classification and extract the symptoms, most of which achieve an excellent performance. In this case, we successfully achieve our research goal. Nevertheless, due to the PATH issue in the PySpark library, we still cannot deploy our NER model onto the chatbot pipeline, which is a pity for our project.

While other researchers focus on predicting COPD morbidity or mortality using indices and statistics from medical experiments, our project conducts the prediction by analyzing the textual data provided by patients themselves. Moreover, while other researchers mainly utilize traditional machine learning methods, our project applies the modern deep learning approach with neural networks. In this case, our project reveals the potential of using complicated ANN structures for COPD symptom classification. Also, it demonstrates the potential in helping patients self-identify if they have COPD-related symptoms, thus urging them to seek a diagnosis in the early stage of the disease. Hence, this chatbot can not only help reduce the cost in traditional health Q&A platforms, but also possibly increase the survival rate of patients in the early stage of COPD.

However, limitations also exist in this project. We do not have access to enough data, such as EHRs or clinic consulting records. We also have a significant unbalance in the data used for the classification task. Both may result in inaccuracy for the final classification performance, though experiments show that ultimately they have little impact on the model performance. Moreover, we currently cannot integrate the NER model into our chatbot pipeline due to the library compatibility issue. For future work, we will try to fix this issue and finish integrating this chatbot pipeline. Also, we want to compare our classification results with more traditional machine learning models like Random Forest and Support Vector Machine. Moreover, we may extend our experiment to other disease categories. In this case, we may develop chatbots that focus on diagnosing various disease types.

6. Conclusion

Our project deals with the problem of COPD symptom recognition. We would like to distinguish COPD symptoms from those of other non-COPD pulmonary diseases, and we construct a machine learning pipeline that detects if the patient has gotten COPD by analyzing textual descriptions and simultaneously extracts corresponding symptoms. We achieve this pipeline by separating the task into a binary classification task, using a BERT classifier, and a NER symptom extraction task, using a pre-trained BertForTokenClassification model. We develop a chatbot that receives descriptions from a patient and outputs the probability that this patient has gotten COPD. This pipeline is still a prototype, but it implies the potential of using the text mining method to help patients conduct self-diagnosis and detect symptoms of specific diseases in the early stage. This can significantly reduce the cost of diagnosis and increase the survival rate of COPD patients.

7. Collaboration

All works are conducted by Xiaoliang (Toby) Zhu alone.

Acknowledgements

The authors appreciate Dr. Zhijun Yin from the Department of Biomedical Informatics at Vanderbilt University for his support and instruction on this project.

Reference

Agarwal, Ankur, Christopher Baechle, Ravi Behara, and Xingquan Zhu. A natural language processing framework for assessing hospital readmissions for patients with COPD. *IEEE journal of biomedical and health informatics*, 22(2):588-596, 2017.

Doğan, Rezarta Islamaj, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1-10, 2014. https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/.

Himes, Blanca E., Yi Dai, Isaac S. Kohane, Scott T. Weiss, and Marco F. Ramoni. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *Journal of the American Medical Informatics Association*, 16(3):371-379, 2009.

Ho, T.W., Tsai, Y.J., Ruan, S.Y., Huang, C.T., Lai, F., Yu, C.J. and HINT Study Group. In-hospital and one-year mortality and their predictors in patients hospitalized for first-ever chronic obstructive pulmonary disease exacerbations: a nationwide population-based study. *PloS one*, 9(12), 2014. doi:10.1371/journal.pone.0114866.

Ju, Meizhi, Andrea D. Short, Paul Thompson, Nawar Diar Bakerly, Georgios Gkoutos, Loukia Tsaprouni and Sophia Ananiadou. Annotating and detecting phenotypic information for chronic obstructive pulmonary disease. *JAMIA Open*, 2(2):261-271, 2019. https://doi.org/10.1093/jamiaopen/ooz009.

Ju, Zeqian, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang and Pengtao Xie. CovidDialog: Medical Dialogue Datasets about COVID-19. *https://github.com/UCSD-AI4H/COVID-Dialogue*, 2020.

Lee, Jung Yeon, Gyu Rak Chon, Chin Kook Rhee, Deog Kyeom Kim, Hyoung Kyu Yoon, Jin Hwa Lee, Kwang Ha Yoo et al. Characteristics of patients with chronic obstructive pulmonary disease at the first visit to a pulmonary medical center in Korea: the Korea COPD subgroup study team cohort. *Journal of Korean medical science*, 31(4):553-560, 2016.

Marin, Jose M., Inmaculada Alfageme, Pere Almagro, Ciro Casanova, Cristobal Esteban, Juan J. Soler-Cataluña, Juan P. De Torres et al. Multicomponent indices to predict survival in COPD: the COCOMICS study. *European Respiratory Journal*, 42(2):323-332, 2013.

Ohar, Jill A., Alireza Sadeghnejad, Deborah A. Meyers, James F. Donohue, and Eugene R. Bleecker. Do symptoms predict COPD in smokers?. *Chest*, 137(6):1345-1353, 2010.

Qaseem, Amir, Timothy J. Wilt, Steven E. Weinberger, Nicola A. Hanania, Gerard Criner, Thys van der Molen, Darcy D. Marciniuk et al. Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the American College of Physicians, American College of Chest Physicians, American Thoracic Society, and European Respiratory Society. *Annals of internal medicine*, 155(3):179-191, 2011.

World Health Organization. Chronic obstructive pulmonary disease (COPD). 2021. https://www.who.int/en/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

Xiong, Wei, Mei Xu, Yunfeng Zhao, Xueling Wu, Bigyan Pudasaini, and Jin-ming Liu. Can we predict the prognosis of COPD with a routine blood test?. *International journal of chronic obstructive pulmonary disease*, 12:615, 2017.