Healthcare Transformation with Informatics and Artificial Intelligence
J. Mantas et al. (Eds.)

doi:10.3233/SHTI230549

Combining NLP and Machine Learning for Differential Diagnosis of COPD Exacerbation Using Emergency Room Data

Fatemeh SHAH-MOHAMMADI[1] and Joseph FINKELSTEIN
The University of Utah, Salt Lake City, UT, USA

Abstract. Chronic Obstructive Pulmonary Disease (COPD) exacerbation exhibits a set of overlapping symptoms with various forms of cardiovascular disease, which makes its early identification challenging. Timely identification of the underlying condition that caused acute admission of COPD patients in the emergency room (ER) may improve patient care and reduce care costs. This study aims to use machine learning combined with natural language processing (NLP) of ER notes to facilitate differential diagnosis in COPD patients admitted to ER. Using unstructured patient information extracted from the notes documented at the very first hours of admission to the hospital, four machine learning models were developed and tested. The random forest model demonstrated the best performance with F1 score of 93%.

Keywords. Chronic obstructive pulmonary disease, machine learning, NLP, differential diagnosis.

1. Introduction

Chronic obstructive pulmonary disease (COPD) is a chronic lung disorder that is expected to become the fifth leading cause of mortality by 2030 [1]. Patients with poorly controlled COPD usually suffer from exacerbations which lead to an accelerated decline in lung function and a reduction of quality of life. During admission to an emergency department, COPD patients initially exhibit a host of symptoms whose presentation overlaps with many comorbidities such as congestive heart failure, coronary artery disease, or atrial fibrillation [2]. Early identification of the underlying cause of exacerbation allows the timely prescription of an optimal treatment plan and helps avoid unnecessary clinical tests and specialist consultations. Since vast amounts of data in EHR are in the form of free text containing clinical findings and symptoms, natural language processing (NLP) has been broadly adopted to (semi-)automatically process free text. The goal of this study was to leverage NLP in the construction of a classification model for differential diagnosis of COPD exacerbation using the clinical notes generated within 24 hours of admission to the hospital. Several previous studies had constructed predictive models for early identification of COPD exacerbation [3-4]. However, none of them incorporated NLP to extract informative features from the notes.

[1] Corresponding Author: Fatemeh Shah-Mohammadi, University of Utah, 421 Wakara Way, Salt Lake City, UT, 84108, USA, E-mail: fatemeh.shah-mohammadi@utah.edu.

2. Method

The dataset was extracted from the Epic electronic health record (EHR) for patients with confirmed diagnosis of COPD. The dataset contained patients' socio-demographic information along with different notes documented within 24 hours of patient admission to ER. The study patients were admitted due to a variety of reasons such as pulmonary embolism, congestive heart failure, COPD exacerbation, and shortness of breath (SOB). A domain expert reviewed discharge diagnoses from discharge summaries to separate the dataset into two classes: admissions definitively resulting from a COPD exacerbation or not. The class labeled as 1 contained the notes for the patients who were discharged with a final diagnosis of COPD exacerbation, while another class (labeled as 0) included the patients' notes also admitted with respiratory distress but discharged due to a condition other than COPD. Table 1 shows the overall composition of the analytical dataset.

To distinguish between the two classes, we considered the very first symptoms and vital signs documented right after admission to the ER, which comprised ER triage notes and ER provider notes. ER provider note contains various information about the patients, including the very first vital signs observed in ER and social determinant of health (SDH) such as smoking, drug and alcohol use status.

To build an NLP pipeline, Clinical Language Annotation, Modeling, and Processing (CLAMP) system [5] was used as an entity extraction tool. Using CLAMP's default named entity recognition toolkit and updating its embedded dictionary, we developed a pipeline to extract the vital signs, i.e., SpO2 level, respiratory rate, pulse rate, and blood pressure (systolic and diastolic pressure) documented in provider notes. Leveraging rule-based components of the CLAMP-GUI, a separate pipeline was also developed to identify the smoking, alcohol, and drug abuse status of the patients. This pipeline developed three tags as follows: "smoking status," "drug status," and "alcohol status." These tags are assigned by constructing a dictionary resulting from a manual review of all provider notes in the dataset and searching for all verbiage and spelling variations of phrases that could be an indication of patients' smoking, alcohol, and drug abuse status. In addition, we developed a dictionary that contained general symptoms of COPD exacerbation and used it to engineer a new feature. Specifically, to account for the very early symptoms documented in triage notes, a new feature was engineered that counts the number of symptoms that appeared in the patient's triage note and checks how many belong to the aforementioned dictionary. This allowed the addition of four highly relevant predictive features to the previously proposed feature set [6]. Moreover, we adapted the dictionaries developed in [6] to the current dataset to add new features. These features incorporate legal, mental, and medical dimensions of distress as well allow accounting for emotional problems, pain and family environment of patients. Patients' age, ethnicity, race, and gender have been extracted as structured data elements from EHR and integrated to the feature set. Patients' sex was labeled as male or female, and race was categorized into White, Black/African Americans, and others.

We compared 4 machine learning models, including support vector machine (SVM), random forest (RF), naïve bayes (NB) and logistic regression (LR). We performed parameter tunings along with 5-fold cross validation to find the best parameters. Cross validation is a robust measure to prevent overfitting. The analytical dataset was randomly divided into 70% training and 30% testing. All analyses were performed in Anaconda Jupyter Notebook, using Python 3.8. The project has been approved by the institutional review board.

Table 1. Composition of the analytical dataset.

Labels	Number of patients (n=134)	%
1 - discharged with diagnosis of COPD exacerbation	60	46%
0 - discharged with a non-COPD diagnosis	74	55%

3. Results

In classes 1 and 0, females account for 72% and 59% with average age of 71 and 70, respectively. Distribution of different races are almost the same for both classes. In both classes 1 and 0, 48% and 41% of patients were former smokers, respectively. The percentage of substance misuse in both classes was almost the same (around 30%). The proportion of patients with alcohol misuse in class 1 was 27%, while it was 15% in class 0. Table 2 presents the most frequent symptoms/conditions detected in triage notes for patients in classes 1 and 0, respectively. The column named "CUI" lists the concept unique identifiers for the symptoms/conditions. The most frequent symptom documented in triage notes for the patient visit in both classes was shortness of breath (SOB). Audible wheezing was the next most common symptom among the patients diagnosed with COPD exacerbation (class 1), and for the patients in class 0, chest pain was the next frequent one. For class 1, mean value for SpO2 level, respiratory rate, pulse rate, systolic and diastolic pressure were respectively 96±2, 19±2, 91±12, 137±20, and 76±12. While for class 0, the mean value for the same measures were respectively 97±2, 18±1, 87±13, 133±21, and 75±12. The mean values for the feature, which is engineered to convey information about mental distress in patients' life [6], in classes 1 and 0 were 13 and 2, respectively.

We developed four predictive models based on naïve bayes, logistic regression, random forest, and SVM machine learning methods (table 3). F1 score and accuracy were used as the evaluation metrics. In SVM, we tuned the regularization parameter (C) and kernel function. The final test set accuracy for SVM was 0.83, with an F1 score of 0.84. In RF, we tuned the number of estimators (trees), and the rest of the parameters were set as default. The best model generated an accuracy and F1 score of 0.93 on test dataset.

4. Discussion

Combining NLP of ER noted with machine learning allowed achieve high accuracy in early differential diagnosis of COPD exacerbation. Extraction of social determinants of health from clinical notes along with the initial presenting symptoms contributed to significant improvement in classification accuracy compared to previous results [7]. Percentage of patients who were former smokers was higher among the patients diagnosed with COPD exacerbation. Higher mean value for the feature associated with mental distress among patients discharged with the diagnosis of COPD exacerbation indicated that mental distress was associated with more frequent exacerbations. These results are congruent with previous reports on COPD exacerbations [8-10]. Among the

tested models, the random forest model showed the highest accuracy with F1 score of 93%.

Table 2. Most frequent symptoms/conditions listed in triage notes stratified by the study classes.

Class 0			Class 1		
CUI	Entity	n	CUI	Entity	n
C0010399	SOB	5	C0010399	SOB	4
C0008031	Chest pain	4	C0043144	Audible wheezing	2
C0010200	Cough	2	C0010200	Cough	2
C0043144	Audible wheezing	1	C0476273	Respiratory distress	2
C0740304	COPD	1	C0740304	COPD	2
C0476273	Respiratory distress	8	C0008031	Chest pain	1
C0278060	Altered mental status	8	C0004096	Asthma	6
C0231835	Tachypneic	6	C0439514	Fever	6
C0018802	CHF	6	C0553668	Labored breathing	4
C0015672	Fatigue	4	C0278060	Altered mental status	4

Table 3. Summary of evaluations.

Models	**Parameters**	**F1-Score**	**Accuracy**
NB	--	0.70	0.73
LR	Alpha (regularization coefficient) = 0	0.78	0.80
SVM	C=100, Kernel= linear	0.84	0.83
RF	Number of estimators = 41	0.93	0.93

5. Conclusion

Combination of NLP-based extraction of unstructured data elements from ER notes with machine learning resulted in significant improvement of classification accuracy. The random forest model was the most accurate machine learning model for early identification of COPD exacerbation in ER.

References

[1] Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS medicine. 2006 Nov 28;3(11):e442.
[2] Pauwels R, Calverley P, et al. COPD exacerbations: the importance of a standard definition. Respiratory medicine. 2004 Feb 1;98(2):99:107.
[3] Fernandez-Granero MA, Sanchez-Morillo D, Leon-Jimenez A. An artificial intelligence approach to early predict symptom-based exacerbations of COPD. Biotechnology & Biotechnological Equipment. 2018 May 4;32(3):778-84.
[4] Boubacar HA, Texereau J. Ensemble machine learning for the early detection of COPD exacerbations. 2017.
[5] Soysal E, et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association. 2018 Mar;25(3).
[6] Shah-Mohammadi F, et al. Using Natural Language Processing of Clinical Notes to Predict Outcomes of Opioid Treatment Program. Annu Int Conf IEEE Eng Med Biol Soc. 2022;2022:4415-4420.
[7] Shah-Mohammadi F, Finkelstein J. Using NLP for Differential Diagnosis of Chronic Obstructive Pulmonary Disease Exacerbation. In 2022 IEEE International Conference on BIBM (pp. 3876-3878).
[8] Kilic H, Kokturk N, Sari G, Cakır M. Do females behave differently in COPD exacerbation? Int J Chron Obstruct Pulmon Dis. 2015 Apr 24;10:823-30.
[9] Finkelstein J, Cha E, Scharf SM. Chronic obstructive pulmonary disease as an independent risk factor for cardiovascular morbidity. Int J Chron Obstruct Pulmon Dis. 2009;4:337-49.
[10] Finkelstein J, Cha E. Association of Veteran Status with COPD Prevalence Stratified By Gender. American Journal of Respiratory and Critical Care Medicine. 2013;187:A6021.