Developing a Rule-Based System for Identifying COVID-19 symptoms
(Assignment 1: BMI 550)

Seyedeh Somayyeh Mousavi
seyedeh.somayyeh.mousavi@emory.edu
Department of Biomedical Informatics, Emory University, Atlanta, GA, USA

Abstract
In response to the global COVID-19 pandemic, this project leveraged NLP to identify a wide spectrum of symptoms associated with the virus. In our methodology, we developed a rule-based system using predefined rules and lexicons and employed both exact and inexact matching approaches. These techniques allowed us to identify symptoms, even in informal language usage. NLP method achieved a Recall of approximately 0.66, a Precision of approximately 0.61, and an F1-Score of approximately 0.63.

Introduction
COVID-19 is a highly contagious viral illness that first emerged in late 2019 and quickly spread globally, leading to a worldwide pandemic [1]. It can manifest with a wide range of symptoms that vary in severity. Natural Language Processing (NLP) plays a crucial role in identifying these symptoms [2]. In the context of COVID-19, NLP is employed to analyze various text data sources, such as medical records, social media posts, and healthcare reports, to detect and identify common symptoms associated with the disease. These symptoms often include fever, cough, shortness of breath, fatigue, loss of taste or smell, sore throat, and body aches [2,3]. NLP enables healthcare professionals and researchers to swiftly identify potential cases, track the virus's spread, and make informed decisions about patient care and public health measures. In this project, we conducted a preliminary study to assess the potential of social media, specifically Reddit's /r/Coronavirus subreddit, as a platform for syndromic surveillance. We analyzed data from Redditors who had tagged themselves as COVID-19 positive using the "flair" attribute in their posts.

Methods
In this project, we initially conducted a manual analysis and annotation of 20 posts focusing on COVID-19-related symptoms, specifically 'Symptom CUIs,' and 'Negation Flag.' We used the COVID-Twitter-Symptom-Lexicon as our reference. During the development of the NLP model, we employed this dictionary for matching and extracting symptoms. Subsequently, utilizing the available data, we created a rule-based system to automatically detect both symptoms and negated symptoms within Reddit posts. A rule-based system represents an approach utilized in Natural Language Processing (NLP), relying on predefined rules or patterns to perform specific tasks, such as text analysis. In the context of detecting concepts or extracting information from text, rule-based systems utilize explicit instructions or patterns to identify and categorize relevant information. This often involves techniques such as exact matching, where the system searches for precise matches of specified keywords or phrases in the text, inexact matching to identify similar terms, and the use of regular expressions to capture complex patterns within the text.

In this project, we performed lowercase preprocessing on the text data. Subsequently, we developed a sequential series of exact and inexact matching approaches to identify symptoms. Exact matching entails searching for predefined words or phrases within the text. Our predefined words are sourced from the COVID-Twitter-Symptom-Lexicon. The following approach utilizes a sliding window method for inexact matching of target expressions within the given text. This process involves dividing the text into sequential windows, with each window's size determined by the length of the expression being sought. The method calculates a similarity score between the content of each window and the target expression. This scoring mechanism accounts for variations in spelling, wording, or slight deviations. When the similarity score surpasses a predefined threshold, the content of the window is recorded for further analysis. The approach also keeps track of the best matching window based on the highest similarity score and reports this best match. This methodology proves valuable in identifying related expressions within text data, especially in scenarios where variations or informal language usage are common, such as in social media content analysis.

In the next stage of our analysis, we employed two distinct methods to determine if a symptom is negated. The first method is associated with symptoms identified during the exact searching process. It involves extracting the text following a negation and tokenizing it into words. The system then checks whether the symptom expression appears within the first three words following the negation. If this condition is met, the function proceeds to search for a period ('.') in the text following the negation. It also examines whether there are any additional negations occurring further in the text. If a period appears after the symptom expression, and there are no other negations occurring before it, the

function classifies the symptom as negated and returns 'True'; otherwise, it returns 'False.' The second approach is linked to symptoms identified through inexact searching. In this process of identifying negations, we tokenize the sentence containing the identified symptoms and search for the presence of negation words. If any negation words are found, the output is 'True'; otherwise, it returns 'False. These two approaches have similar results, so we merged them to have access to the Unix symptoms.

Results

The first part of this assignment involved manual annotation. Table 1 displays the Inter-Annotator Agreement (IAA) scores between my annotations and those from other annotators. We achieved a Recall of approximately 0.6584, a Precision of approximately 0.6073, and an F1-Score of approximately 0.6318 for our NLP method. Figure 1 demonstrates the distribution of ICU symptoms. In total, we have 45 labels, with an additional category labeled 'other'. 'Pyrexia,' 'Body ache & Pain,' and 'Cough' are the most frequently occurring symptoms.

However, this model exhibits several limitations that impact its performance. Despite comprising both exact and inexact components, our model heavily relies on predefined word lists and dictionaries to identify specific terms or concepts within the text. This approach may not effectively capture all relevant instances, especially in the context of social media data. Consequently, the method may overlook a significant number of expressions or instances related to the target concepts simply because they are not present in the lexicon or because they appear in nonstandard forms, such as misspelled words or slang. For instance, terms like 'pain' can manifest in various contexts and associate with numerous other words, making precise matching with ICU symptoms challenging. Additionally, while we employ fuzzy thresholding, determining the optimal threshold remains a complex task.

Table 1. IAA scores between S15 annotation file and the annotations from other annotators

Annotators	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S15
IAA score	0	1	0.76	0.94	0.83	0.86	1	0.88	0.82	0.81	1	0.93	0.82	0.83

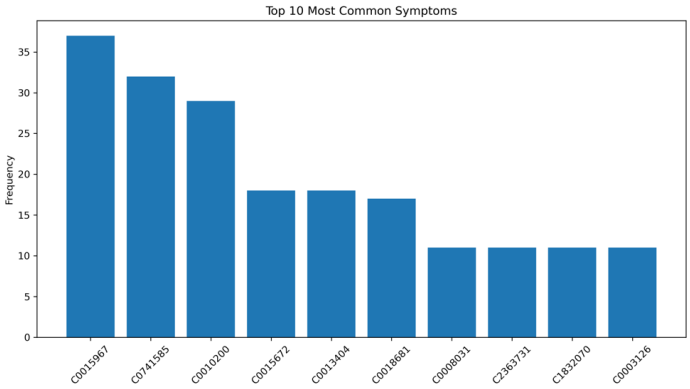

Figure 1. Distribution of the ten most frequently identified symptom groups by the NLP model.

Link to materials

1. Code
2. Annotated file (S14)
3. Annotated unlabeled file
4. Annotated gold standard file

References

1. Cascella, Marco, Michael Rajnik, Abdul Aleem, Scott C. Dulebohn, and Raffaela Di Napoli. "Features, evaluation, and treatment of coronavirus (COVID-19)." (2020).
2. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang YC. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. J Am Med Inform Assoc. 2020 Aug 1;27(8):1310-1315. doi: 10.1093/jamia/ocaa116. PMID: 32620975; PMCID: PMC7337747
3. Amin, Md Tanzilul, Mahmud Hasan, and NM Mahmudul Alam Bhuiya. "Prevalence of covid-19 associated symptoms, their onset and duration, and variations among different groups of patients in bangladesh." Frontiers in public health 9 (2021): 738352.