A Preliminary Study on COVID-19 Symptom Detection Using Social Media Data with Fuzzy Matching Technique

Masoud Nateghi, BSc[1]

[1]Emory University, Atlanta, Georgia, United States

Abstract

In this research, we conducted a preliminary examination of COVID-19 symptom detection by employing a fuzzy matching technique in conjunction with a straightforward negation detection method. We designed an automated approach to establish the optimal threshold for fuzzy matching process, guided by an error analysis of the rule-based system. This approach yielded promising results with a recall of 0.663, precision of 0.757, and an F1-score of 0.707.

Introduction

COVID-19 stands as one of the most devastating pandemics in human history. Since the onset of the pandemic in early 2020, it has claimed the lives of over 6.6 million individuals worldwide [1,2]. Previous research has predominantly focused on examining the symptoms of severely affected individuals requiring hospitalization [2]. However, the early detection of symptoms carries significant importance for affected individuals, as it can help curtail the spread of the disease by raising awareness of their condition. Additionally, early symptom detection can lead to substantial cost savings in treatment, as patients can receive timely medical intervention during the initial stages of the disease. This study advocates the potential use of social media data and natural language processing techniques for the early identification of symptoms.

Materials and Methods

We selected Reddit as our data source and gathered information from the subreddit /r/Coronavirus, where users employ the flair attribute to signify their positive COVID-19 test results.

1. Annotation and Agreement

We manually annotated a set of 23 Reddit posts to gain insights into the problem at hand. This annotation task was conducted independently by a group of individuals to assess the agreement among humans, as the ability to reach consensus in human annotation is crucial for the effectiveness of machine-based annotation. Table 1 illustrates the inter-annotator agreement between the author and other annotators using Cohen's kappa metric. This analysis supports the notion that automatic annotation may be a feasible objective.

Table 1. Inter-annotator agreement of the author (s9) with other annotators using Cohen's kappa metric.

annotators	s3	s4	s6	s7	s10	s11	s12	s13	s14	s15
s9 (author)	0.865	0.928	1	0.938	0.860	0.920	1	1	0.171	0.803
common posts	7	10	1	2	21	2	1	1	4	8

The selected posts all exhibited at least one COVID-19 symptom. Annotators also incorporated the essential concept of negation detection, where patients explicitly indicated the absence of common and prevalent symptoms. In this scheme, a value of ‘1’ indicated negation, while ‘0’ indicated the presence of a specific symptom. Additionally, for a more comprehensive analysis of various symptoms, annotators mapped each symptom expression to a standardized and unified concept ID and expression. For instance, different expressions used by individuals to convey having a fever, such as ‘my body temperature is high’ or ‘I have a slight fever’ were all mapped to a singular concept of Pyrexia with the unique ID C0015967.

We extended our evaluation to include a set of 35 unseen posts to assess the designed system's capability to accurately identify true symptoms while avoiding false ones.

2. System Design

Initially, we conducted a brief preprocessing step on the unseen posts and the lexicon file by converting all text to lowercase. This preprocessing aimed to mitigate conflicts arising from differences in letter case.

As one of our inputs, we utilized the ‘covid-twitter-symptom-lexicon’ file. Our system was designed to begin by storing all symptom expressions in a dictionary data structure. It then iterated through each symptom expression, employing a sliding window with a size equal to the number of terms present in the symptom expression. Subsequently, the system calculated the Levenshtein distance, a metric based on the idea that words with similar characters have a smaller distance. To refine this distance measurement, we applied the Levenshtein ratio, producing values between zero and one. Typically, strings that exhibited some dissimilarity, often due to misspellings or conceptual proximity, yielded higher Levenshtein ratios.

The use of the Levenshtein ratio led us to consider employing stricter (higher) thresholds for single-term expressions such as ‘fever’ or ‘cough’ where we expected less dissimilarity between the expressions in the posts and those in the lexicon file. Conversely, milder thresholds seemed appropriate for longer expressions like ‘temperature will randomly go up’ and ‘lungs feel like they are burning’. As a result, we established the threshold as 1 - 0.7 * length(expression), and this threshold cannot go below 0.8, so we used a minimum function to restrict the threshold. This choice aimed to minimize the capture of minor textual similarities among symptoms and reduce the risk of detecting similar expressions for distinct symptoms. For example, using a constant threshold (0.75) for all symptoms might lead to identifying ‘ever’ as ‘fever’ or interpreting ‘lost my sense of smell’ as both ‘loss of taste’ and ‘loss of smell’.

To address the challenge of identifying negated expressions, we devised a straightforward negation detection method. Once we had identified all the expressions using inexact matching, we examined each expression in the context of its respective sentence. Specifically, we checked for the presence of common negation words such as ‘no,’ ‘not,’ ‘did not,’ ‘with no,’ and etc. If a negation word was detected in the sentence, we extended our analysis up to four words ahead to determine whether the symptom was negated. This assessment was guided by two simple rules.

The first rule was applied when another negation word immediately followed the initial negation word, as in the phrase ‘I did not have no fever.’ In such cases, the symptom following the second negation was considered as not negated.

The second rule came into play when a negation word was followed by a period. In this scenario, the symptom located between the negation and the period was considered negated. However, it's important to note that symptoms appearing after the period were not necessarily negated. For instance, in the sentence ‘I had no fever or cough.’ both ‘fever’ and ‘cough’ were interpreted as negated. Conversely, in the sentence ‘I had no cough. Fever only.’ the symptom ‘cough’ was identified as negated, while ‘fever’ was not.

Results

To assess the performance of the developed system, we employed a set of 35 unseen Reddit posts, utilizing metrics such as recall, which measures the system's accuracy in detecting true symptoms; precision, which gauges the proportion of correctly identified symptoms among the total detections; and F1-score, which offers a combined evaluation of both metrics. Table 2 presents the system's performance based on the aforementioned metrics.

Table 2. F1-score, Recall, and Precision for the designed system.

Metric	F1-score	Recall	Precision
Our Method	0.707	0.663	0.757

As illustrated in Figure 1, by associating each expression with the corresponding standard symptom using the 'covid-twitter-symptom-lexicon' file, valuable insights into the early common symptoms of COVID-19 can be obtained.

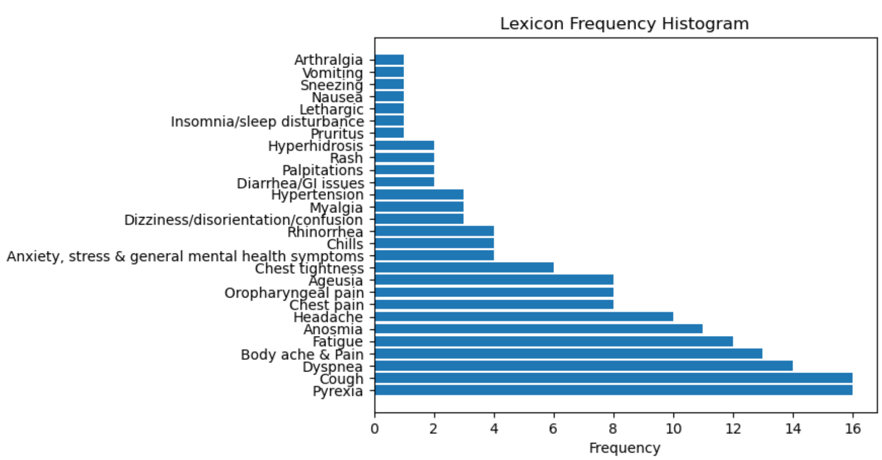

Figure1. Distribution of early common symptoms of COVID-19

Error Analysis

1. System-driven errors

Further analysis of the system's output reveals that it performs quite well for symptoms that have limited lexical variation, such as ‘fever’ or ‘cough.’ However, as the number of expressions and ways to describe a concept or symptom increases, the system begins to show some weaknesses. For instance, ‘chest pain’ can be expressed in numerous ways, and manually annotating all possible expressions for each symptom is impractical and time-consuming. This limitation arises from the nature of inexact matching, which requires a comprehensive lexicon to excel.

Another challenge arises with negation, where individuals might express a symptom but later indicate that it has resolved. For example, someone might say, ‘I had a fever on the first day, but I don't feel any fever now.’ The system tends to recognize the last expression as the best match, resulting in a false negation. This issue can be mitigated by stopping the search for additional occurrences of the same expression once the first one is found.

2. Lexicon-driven errors

Additional errors stem from the lack of diversity in expressions for certain symptoms, such as ‘rash’ or ‘heart’ (an organ affected by the virus). The system struggles to detect abbreviations like ‘sob’ or ‘gi’ which adding them to lexicon file can lead to false positives like detecting ‘so’ for ‘sob.’ Moreover, expressions like ‘sore-C0741585’ or ‘pain-C0741585’ are associated with ‘Body pain,’ and the system mistakenly predicts ‘body pain’ symptoms whenever it encounters ‘pain’ or ‘sore,’ even when the true symptoms may not be related to body pain, such as ‘sore throat.’ To address this, it may be advisable to exclude such overly general terms applicable to multiple symptoms.

Lastly, the system struggles to detect uncommon or exotic symptoms like ‘blocked ears’ due to the absence of corresponding expressions in our lexicon file. This highlights the need for a system capable of automatically identifying phrase-like symptom expressions, a task that rule-based systems cannot fully achieve.

Discussion and Conclusion

Social media data is renowned for its inherent noise, primarily characterized by the informal expressions that people frequently employ in their daily discussions and opinions. Despite the complexity of natural language processing in the context of Biomedical NLP forced by its restricted domain, our study strongly suggests that we can extract early COVID-19 symptoms from social media data. As illustrated in Figure 1, it becomes evident that symptoms such as pyrexia, cough, dyspnea, body ache, and fatigue emerge as the most prevalent symptoms across all the symptom expressions expressed by Reddit users. This observation aligns with our initial expectations.

Supplementary Materials

All the files and material used in this project can be found here:

References

1. Lui G, Guaraldi G. Drug treatment of COVID-19 infection. Curr Opin Pulm Med. 2023 May 1;29(3):174-183. doi: 10.1097/MCP.0000000000000953. Epub 2023 Mar 14. PMID: 36917228; PMCID: PMC10090306.
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.