Enhanced Identification of COVID-19 Symptoms Through a Rule-Based System Applied to Social Media: A Concise Analysis

Alireza Rafiei, MS[1]
[1]Department of Computer Science and Informatics, Emory University, Atlanta, GA

Abstract

The landscape of public health surveillance is undergoing a transformation with emerging avenues like social media, which provide new possibilities for monitoring infectious diseases. This concise analysis presents a preliminary investigation into the feasibility of using the Reddit platform for surveillance of COVID-19 symptoms. The study employs annotated datasets of Reddit posts for the system development and a gold standard dataset of 34 unseen posts for evaluation. We introduce a robust data preprocessing pipeline to prepare Reddit posts for analysis and design a rule-based natural language processing system that identifies symptoms and their negation flags. The system is able to achieve an f1 score of 0.685 on the evaluation dataset. The results of this investigation provide valuable insights into the limitations and advantages of leveraging social media data for public health surveillance, offering a new paradigm for the timely and comprehensive monitoring of infectious diseases like COVID-19.

Introduction

In recent years, syndromic surveillance has played an increasingly pivotal role in public health initiatives, especially for the timely identification and management of infectious diseases. Traditional methods often rely on hospital or clinical data, which predominantly includes information from those who are severely ill, thereby potentially missing the broader trends that could be observed in the general population. This limitation, along with other challenges such as delayed reporting, limited geographic coverage, and the potential for underrepresentation of marginalized communities, led a team of public health professionals to explore alternative avenues for collecting syndromic information [1]. They approached us with a unique question: "Could social media platforms serve as a viable medium for conducting syndromic surveillance to track COVID-19 symptoms?"

This concise analysis presents a preliminary study designed to answer this intriguing query. We chose Reddit, a popular social media platform, as our primary data source for this investigation. Reddit enables users to engage in topic-specific discussions within various communities known as "subreddits." One such subreddit, /r/Coronavirus, proved to be a rich resource for our study. Notably, users can mark their posts with a "flair" attribute to indicate that they have tested positive for COVID-19, thereby providing a potentially invaluable data set to analyze symptoms as reported by a broad section of the population. This dataset allows us not only to gain insight into the range of symptoms experienced by COVID-19-positive individuals but also to gauge the temporal evolution of these symptoms. As a result, we focused on assessing the feasibility of utilizing social media data for syndromic surveillance using a rule-based system to explore the limitations and advantages of such an approach. The outcomes of this study aim to inform public health professionals by offering a more comprehensive and timely understanding of infectious diseases like COVID-19.

Materials and Methods

Dataset. We have annotated a set of 22 random Reddit posts that contain at least one mention of COVID-19 symptoms or more. The objective of this annotation process is to extract various types of information, including the expression of the symptom, the standardized symptom term, the symptom's Concept Unique Identifier (CUI) based on a specialized COVID-19 lexicon, and a negation flag [2]. The negation flag indicates whether a Reddit user has explicitly stated that they do or do not have the symptom in question. This annotated dataset serves two primary purposes: first, it provides valuable insights into the structural nuances of the problem at hand, and second, it helps in the development of a rule-based system for symptom detection. To evaluate the efficacy of this developed system, we subsequently use a "gold standard" dataset composed of 34 unseen random Reddit posts.

Preprocessing. We design a data preprocessing pipeline for accurate and effective downstream analytics. The initial step of the pipeline involves converting the entire posts to lowercase. This is followed by the elimination of specific punctuations including periods, commas, semicolons, colons, and certain bracketing symbols, among others. The rationale behind this is to reduce lexical noise and improve the focus on meaningful terms, especially for matching the terms extracted from sliding windows. Utilizing regular expressions, the pipeline is configured to identify and remove any date patterns that follow conventional formats. Moreover, it prunes redundant or "stop" words like 'of', 'in', and 'to', which although grammatically essential, often carry minimal semantic weight and can introduce noise in text analytics processes. Lastly, the pipeline addresses the issue of ambiguous or unclear descriptions by replacing phrases like "smell/taste" with a more descriptive and unambiguous term for a system, such as "smell* and taste*". As a result, if we see phrases like "loss of sense of smell/taste", the system can identify both "loss of sense of smell" and "loss of sense of taste" symptoms. The output is a cleaned post ready for downstream natural language processing systems.

System. We have designed and analyzed a rule-based natural language processing system that is capable of detecting symptoms of COVID-19 from Reddit posts. The system works based on the concept of inexact matching. Of note, Levenshtein string similarity is employed to recognize symptoms while also accounting for near-misspellings and paraphrased expressions. Upon tokenizing a post, the system uses a rolling sliding window that traverses the text with varying window sizes, ranging from 1 to 9, with a stride of one. This approach effectively captures both single-word and multi-word symptoms. Of note, the system is engineered to prevent the redundant detection of the same symptom using different window sizes. In addition to identifying symptoms, the system incorporates a feature for negation detection. To accomplish this, we use a preprocessed list of negation triggers, which are precisely matched with the text to fit seamlessly into the analysis pipeline. Negation is established based on context; whether a symptom mentioned appears within the scope of any preceding negation phrases with the consideration of terms following the negation. If a negation is detected, it is flagged accordingly in the final result.

Results

The performance metrics of the developed system are assessed using a gold standard dataset. In this regard, the system is designed to recognize symptoms as well as their corresponding negation flags. Utilizing the labels provided through manual annotation of the evaluation dataset, key metrics such as recall (number of true positives/all actual positives), precision (number of true positives/all predicted positives), and the F1 score (the harmonic mean of precision and recall) are calculated and reported in Table 1.

Table 1. Different performance metrics of the proposed rule-based system evaluated on the gold standard dataset.

Performance Metrics	**Recall**	**Precision**	**F1 Score**
Proposed System	0.757	0.624	0.685

Figure 1 shows the word cloud of the detected COVID-19 standard symptoms in the test dataset using the developed system. Pyrexia and body ache & pain were the most frequently detected symptoms by the system.

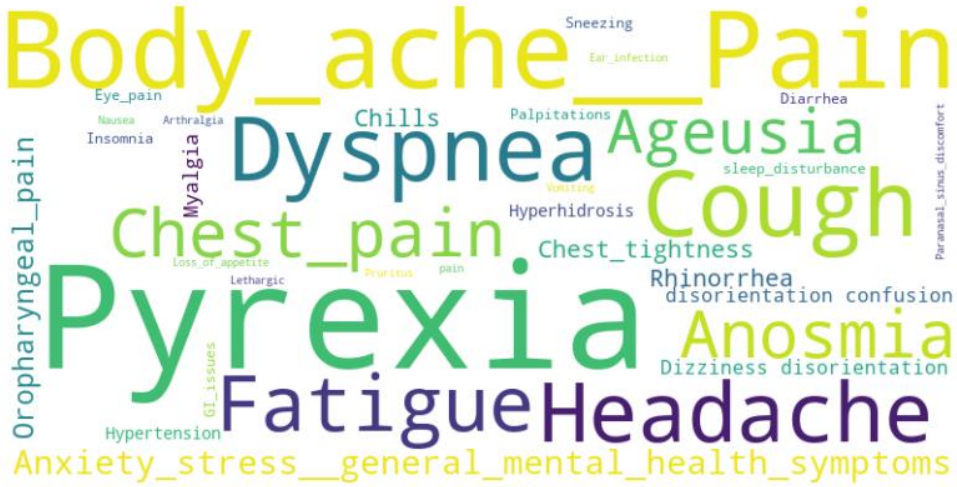

Figure 1. Word cloud of detected COVID-19 standard symptoms in the gold standard dataset using the proposed rule-based system. The size of each term reflects its frequency of mention among Redditors. Multi-word symptoms like 'shortness_of_breath' are represented as single terms for clarity.

Limitation

The analysis presents multiple limitations at both the system and data levels. On the system side, the primary constraint is its heavy reliance on a predefined lexicon for identifying COVID-19 symptoms. Consequently, if a symptom description is absent from the lexicon, the system will fail to report it. As for data-level limitations, there are two major issues. First, the number of posts available for testing the system is relatively limited, which may affect the robustness of our evaluation. Second, the manually annotated gold standard dataset contains several errors, undermining the system's performance in reporting accurate results.

Conclusion

The findings of this analysis highlight the potential for integrating social media platforms like Reddit into traditional public health surveillance methods. Our rule-based system, designed to identify and analyze mentions of COVID-19 symptoms in Reddit posts, demonstrates promising results in terms of recall, precision, and F1 score. The dual capability of symptom detection and negation handling allows the system to deliver nuanced insights into public sentiment regarding COVID-19 symptoms.

Availability of data and materials

All the materials and scripts used in this study can be found here: https://github.com/AlirezaRafiei9/BMI-550/tree/main/Assignments/Assignment%201

References

1. Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. Journal of the American Medical Informatics Association 2004;**11**(2):141-50

2. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang Y-C. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. Journal of the American Medical Informatics Association 2020;**27**(8):1310-15