Rule-Based System for Automated Identification of COVID-19 Symptoms

Avinash Kumar Pandey (Finance PhD Student)
Emory University, USA

Abstract

This report presents a rule-based system for the automated identification of COVID-19 symptoms. The system utilizes a combination of exact matching, regular expressions, fuzzy matching with multiple windows, and associated negation flag detection. The system achieves a F1 score of 0.64 when benchmarked against a Gold Standard Evaluation excel file with pre-annotated labels. This report is submitted as part of Assignment_1 of BMI 550 that required us to build a rule-based system that automatically detect Covid-19 symptoms. Additionally, the output file containing rule based annotated symptoms of UnlabeledSet Excel file is also submitted.

Methodology

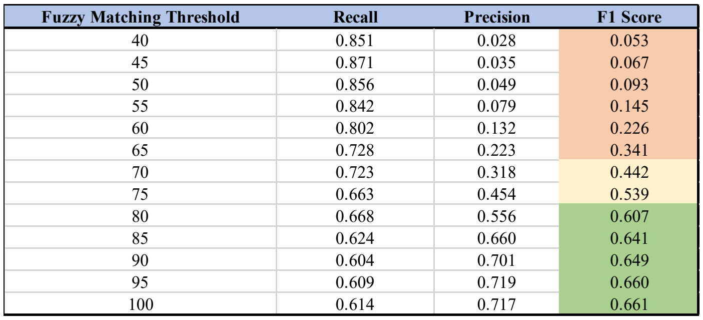
The rule-based symptom identifier is designed to identify Covid-19 symptoms mentioned in a set of Reddit posts. The system first preprocesses the Reddit post text by lower casing and removing "/n" & "*/n" terms and then tokenize it for further analysis.

The system employs three different types of matching techniques. First, it employs regular expression (regex) matching to find exact matches of symptoms in the post. For each symptom in the provided list, it constructs a regex pattern using re.compile and searches for matches in the Reddit post using *re.escape()* function. If a match is found, the symptom is added in matched symptoms lists. Next, the system applies fuzzy matching with a *rolling window of size 2 and 7* using *fuzz.ratio* function which computes the *levenshtein ratio,* which quantifies the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another . If the similarity score is above the specified threshold, the symptoms is added to the matched symptoms list. Finally, any duplicate entries in the matched symptoms are removed, ensuring a unique set of matched symptoms and corresponding text fragments. The system also checks for the presence of negation terms for each identified symptom and raises the negation flag if negation is detected. Results are published following Mandl et al[1] and Sarker et al.[2]

IAA Agreement :

The Inter Annotator Agreement (IAA) assessment between manual symptom annotations suggests high level of agreement, indicating consistent and quality annotations. Instances of slightly lower agreement can be attributed to individual judgements of posts and the false inclusion of symptoms related to personal medical history, as well as those of family and friends. Despite these variations, the overall IAA underscores the robustness of the manual annotation process. The average IAA value across of s7 excel file with other students is **0.94** and there are 8 instances of complete agreements with other submissions

Results : Using the combination of regex and rolling window fuzzy matching with 2 & 7 terms window, F1 scores at various thresholds were computed. The Table 1 below shows the F1 scores, precision and recalls at thresholds ranging from 40-100. The last row can be interpreted as the combination of regex and exact matching with rolling windows. The performance of the system plateaus above the threshold value with F1 score increasing marginally from 0.64 to 0.66 with maximum value occurring at threshold of 100 with precision value of 0.72.



Fuzzy Matching Threshold	Recall	Precision	F1 Score
40	0.851	0.028	0.053
45	0.871	0.035	0.067
50	0.856	0.049	0.093
55	0.842	0.079	0.145
60	0.802	0.132	0.226
65	0.728	0.223	0.341
70	0.723	0.318	0.442
75	0.663	0.454	0.539
80	0.668	0.556	0.607
85	0.624	0.660	0.641
90	0.604	0.701	0.649
95	0.609	0.719	0.660
100	0.614	0.717	0.661

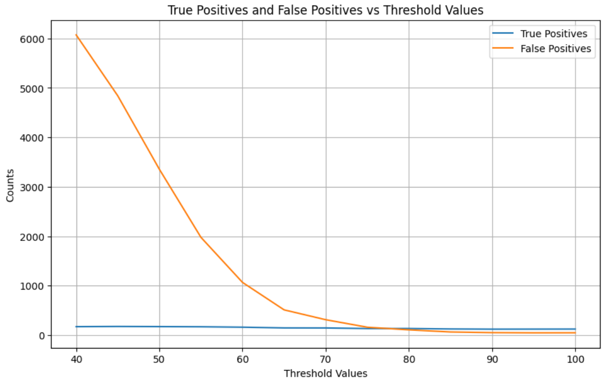
Table 1: Multiple Window Fuzzy Match + Regex Model Performance across different threshold values

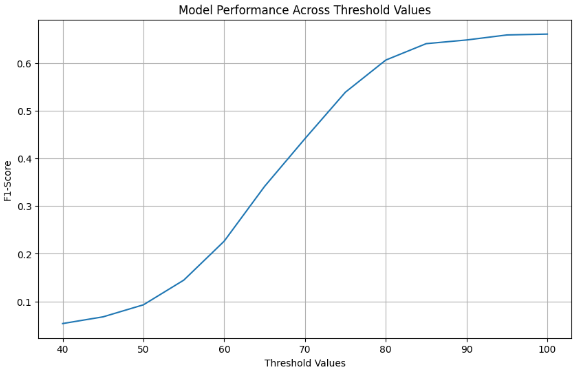
Important Links: Overall Github repository link is here. Automated Unlabeled Annotated File Link is here.

Error Analysis:

Several notable categories of errors have been identified, including fuzzy matching inaccuracies resulting in false positives, missed negations, failure to identify symptoms in compound symptom expressions, non-standard symptom description by social media users, spelling errors, and potential data inconsistencies/error in the Gold Standard File.

One prominent category of error involves *fuzzy matching inaccuracies*, where similar-sounding expressions are incorrectly identified as symptoms. For instance, the system erroneously matches "feeling pain" and "feeling good," “breathing fine” matches “breathing pain”, *“leg pain” matches “lung pain”.* Increasing the fuzzy threshold reduces the possibility of these errors. To illustrate this effect, the plot of number of false positives and fuzzy threshold is shown :





The system also has issues in *capturing complex negations* in instances like "cough gone away" being incorrectly identified as a symptom. To mitigate this, a context-based model or expanded negation list is recommended to improve the performance. The system also struggles to identify symptoms in *compound symptom expressions*, such as "I lost my sense of taste and smell” is only matched to *“I lost my sense of taste” and not the “lost my sense of smell*” even with wider window values. Reducing the threshold value of fuzzy matching increases the false positives. Instances where users employ *non-standard complex terms to describe symptoms, such as "shitty breathing" or "I am wiped out,"* pose a challenge to the system. A more comprehensive lexical file that includes a broader range expressions and variations can mitigate this issue. Finally, *spelling errors* in symptom descriptions can lead to missed matches in exact matching. Implementing a *spell-checking mechanism* may help address this issue for social media texts. Finally, there are some inconsistencies in the Gold Standard File, such as the presence of only two "$$" signs in rows 19 and 20, impacting the F1 score. In the Unlabeled excel file also, there was an empty row in row 275-ga7gvk user may cause error in system codes of various students.

Conclusion/Discussion:

In conclusion, the system used a combination of regex and fuzzy matching over different rolling windows. The system achieved a F1-score of 0.64 and precision value of 0.717. One important finding of this exercise is the abnormally high values of false positives for lower value of thresholds for fuzzy matching. In the error analysis , we also saw that a system that could use contextual annotation could outperform the rule based model. Also, having a more comprehensive negation list, spelling checks on posts and lexical dictionary could also improve performance. However, the system could be used for initial auto annotation for large number of posts at scale and thus is a significant tool with good potential to contribute to BMI research.

References

1. Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc. 2004;11(2):141-150. doi:10.1197/jamia.M1356.
2. Sarker A, Lakamana S, et al. 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. ;PMCID: PMC7337747.