COVID-19 Syndromic Surveillance from Reddit Comments

Sohail Nizam
Emory University, Atlanta, Ga

Resources

All files associated with this project (including code and data) can be found at the following location: https://github.com/SohailNizam/BIO-NLP-Assignment-1

1 Methods

1.1 Inter-Annotator Agreement

12 people were each given a set of Reddit comments to annotate. For a given comment, an annotator would identify symptoms mentioned and record the exact phrasing used by the person who wrote the comment, the standard medical expression used to denote such a symptom, the corresponding CUI, and a flag to indicate whether the symptom was negated or not. Cohen's Kappa was then calculated to measure Inter-Annotator Agreement (IAA) between each person's annotations and the annotations of the same set of comments from other people. Table 1 shows the results of the IAA analysis.

Annotator	1	2	3	4	5	6	7	8	9	10	11	12
Kappa	.86	.81	.99	.96	.90	.85	1.00	.84	.81	.78	.82	.84

Table 1: *Cohen's Kappa measure of Inter-Annotator Agreement for each annotator. Scores are based on varying numbers of comments.*

Such variation is expected as seen in Table 1 is expected. These scores indicate high agreement between annotators. This suggests that symptoms in these Reddit comments are identifiable and that there is potential for an automated system to do well with annotation.

1.2 Exact and Inexact Matching

The system was built to identify both exact and inexact matches with symptoms in the lexicon. Each comment was tokenized into sentences. Then, each sentence was searched by moving windows of varying sizes across the words and evaluating whether or not each window was found in the symptom lexicon. All window sizes from a single word to the entire length of the sentence were checked. At a given window, if the words did not exactly match any symptom expression in the lexicon, Levenshtein Ratios were computed between the words in the window and each symptom expression in the lexicon. The symptom expression with the highest similarity to the words in the window was saved, and the Levenshtein Ratio for this symptom was compared to a pre-set threshold value which we consider to be a tuning parameter. If the Levenshtein Ratio was found to be higher than the threshold, the words in the window were considered to be a match for the symptom. Threshold value choices are discussed in detail in the Results section. To prevent single words being identified in multiple unique symptoms, window sizes were iterated through from largest to smallest, and whenever a symptom was detected, flags for each word in the window would be stored indicating that they had already been used in finding a match.

1.3 Negation

Any time a new symptom match was identified, the sentence was checked to see if there was an occurrence of a negation word or phrase appearing before it up to three words away. One common occurrence in the

comments was for a symptom to be mentioned first in the affirmative and then later in the negative. To account for this, if a symptom was ever recorded as both positive and negative, the negative symptom record was removed. A feature was also added to only consider a window to be a match if it did not contain the word 'no' thus avoiding matching 'no fever' with 'fever', for example.

2 Results

The system was tuned on a set of 20 human annotated comments. The symptoms found in these comments during annotation along with the originally provided lexicon together made up the full symptom lexicon. Success of the system was quantified by the F1-score primarily and by precision and recall secondarily. Table 2 shows results for the system evaluated at several different Levenshtein Ratio threshold values.

Threshold	F1 Score	Recall	Precision
.80	.865	.870	.860
.85	.862	.848	.876
.90	.890	.837	.951
.95	.877	.815	.949

Table 2: *Results for the annotation system evaluated on the author's personally hand annotated comment set. Threshold indicates the Levenstein Ratio threshold beyond which inexact matches were considered to be true matches.*

Table 2 indicates that the Levenshtein Ratio threshold of .90 yields the best results when using F1-score as the main metric for evaluation. With this threshold chosen, the system was evaluated on the already annotated Gold Standard set. The resulting F1-score was .752, the recall was .665, and the precision was .865. The drop in quality from the hand annotated set to the Gold Standard set is to be expected as all of the symptoms written in the hand annotated set were added exactly to the symptom lexicon. The fact that the F1 score did not drop drastically indicates that the system has done well in generalizing to a new test set.

3 Errors and System Limitations

There were several limitations to this annotation system. One issue is that the negation symptom is quite simple. It is unable to pick up negations such as in the phrase "it wasn't accompanied by any other symptoms like shortness of breath." Even with the addition of "wasn't accompanied by" to the list of negations, it is still too far from the actual symptom to detect. Furthermore, the system has trouble with lists of negations such as "no fever, chills, headache, or shortness of breath." The longer the list gets, the more symptoms will fall outside of the set three word scope of the negation. Future work could seek to deal with list negations specifically. Finally, even with a finely tuned threshold, inexact matching based on Levenshtein Ratios can yield surprising results, finding matches between phrases that would not be considered similar by a human.

4 Conclusion

The results of the evaluations show that this system could be used as a preliminary measure to automatically detect mentions of symptoms in social media posts. The varying results based on Levenshtein Ratio threshold choices indicate that this threshold could be changed depending on the goal of the user. For example, the threshold could be lowered if the user is interested in capturing all possible mentions of symptoms and is not concerned with false positives. The threshold could be raised if the user is only interested in exact matches for a set list of symptoms. There is room for improvement. However, even as it is this system would serve as a good precursor to one built on much larger quantities of data using deep learning techniques.