Sunday, August 27, 2023

Assignment 1: BMI 550— Applied BioNLP

This assignment is based on a real-world scenario

Version: 1.3 *(draft; further details will be added regarding submission protocols)*

Background

You are a data scientist and have been approached by a team of public health professionals. The team specializes in *syndromic surveillance* and want to track symptoms of COVID that are reported over social media. They explain, "*We have been collecting data from emergency departments, but all that data is from patients who are seriously ill. How can we track symptoms experienced by the general population? Is it possible to obtain syndromic information from social media?*"

Your responsibility now is to design and execute a *preliminary study* to ascertain if social media is a possible platform for conducting syndromic surveillance. You have chosen Reddit as your data source and have collected data from the subreddit /r/Coronavirus. Within that subreddit, *Redditors* can use the *flair* attribute to indicate that they have tested positive for COVID19, and you have already collected their posts. Now you have to carry out a set of steps to execute the preliminary study.

Step 1.

You will manually analyze and annotate some data to identify the COVID19-related symptoms mentioned. Specifically, you will tag every (i) symptom and (ii) negated symptom from at least 15 Reddit posts. Further details are provided in the *Specific Tasks* section.

Step 2.

You will compute inter-annotator agreements (IAA) between your annotations and the annotations performed by your classmates (see IAA_Calculator.py). The IAA will provide us an idea of how difficult (or easy) the task of symptom annotation is for humans. If agreements are relatively high, it may be possible to develop an automated system for the task. Otherwise, it is important to understand why and where humans tend to disagree. Based on this analysis, *annotation guidelines* can be prepared. Such annotation guidelines also help to ensure that your research is reproducible (and, hence, useful in real life).

Step 3.

Using the annotated data, you will develop a system that attempts to automatically detect symptoms and negated symptoms from Reddit posts. Ideally, the annotated data should be used to train a sequence labeling system to detect symptoms automatically. Conditional random fields have been very popular for many years, and recently variants of recurrent neural networks have produced state-of-the-art performances in sequence labeling tasks. They, however, require large amounts of annotated data. Since this is a preliminary study, you choose not to apply these

machine learning algorithms. Instead, you will build a rule-based system for the symptom detection task. For the development, you will use the data annotated by yourself and others. 80% of all the annotated data will be made available to the students for this assignment. The remaining 20% will be used for evaluation.

Step 4.

You will perform an evaluation and *analysis of errors,* and write a brief report (2 pages max.) *outlining* your methods, results, and conclusions.

Specific Tasks

Annotation.

You are provided with an excel file with Reddit posts and a dictionary for the symptoms we are planning to monitor (any other symptom should be labeled as 'other'). The excel file will contain a sample row annotated. You have to complete at least 15 more. You have to fill up 4 columns: in the *Symptom Expressions* column, copy-paste each symptom expression exactly as mentioned by the user, and separate each symptom using the $$$ marker. In the *Standard Symptom* column, put the standard expression for the symptom from the symptom dictionary. In the *Symptom CUI* column, put the CUI for each symptom. Put all symptoms in the same order in different columns and always separate by using $$$. The symptoms and negated symptoms will be in the same column. A fourth column, *Negation Flag*, is where you will put a series of 1s and 0s to indicate if a concept in the previous column is a negated symptom or not. See the example in the annotation file.

Each student will be provided a unique annotation file with 20-25 posts. You can choose 15 posts to annotate (the least amount required). You are, of course, welcome to annotate more, but it will not result in higher grade. It may, however, help improve your system.

Annotation Agreement

Using a script provided (**IAA_Calculator.py**), you will compute IAA between your annotations and the annotations from other annotators. A subset of your posts will be annotated by some of the other students and you will not be aware of which the common posts are before the annotations are complete. You will briefly analyze several posts with annotator disagreements, correct any clear error, and note any aspect where the correct annotation was not clear. Do you think there is good agreement amongst the annotators? You will describe this in a brief paragraph in your report.

Rule-based System

Typically, such rule-based systems apply methods such as exact matching, inexact matching, regular expressions, and so on, for detecting concepts. These are mostly lexicon-reliant approaches, where annotated data provides *supervision* for finding standard and non-standard expressions. Since this is social media data, we can expect to

find diverse expressions of symptoms. Given a set of Reddit posts, the system will output a text file that contains, tab separated, the *id* of a post, the symptom expression (or entire negated symptom expression), the *CUI* for the symptom, and a flag indicating negation (*i.e.*, **1** if the symptom expression is a negated one, **0** otherwise). This text file will be used for evaluation of the final system. It is important that your output is exactly as the sample provided. Otherwise, the evaluation system will not be able to read your file properly.

Evaluation, Results and Discussion
Your system will be evaluated using the F_1-score metric. You will be provided with a small number of additional annotated files in week 5, and an evaluation script, which you will use for evaluation of your system. In your report, discuss the errors your system made and how it could have been improved. You will also be provided with a set of unannotated files (*i.e.*, containing ID, date and text only), and you will also provide a link to the output of your system for this file.

Your report should be accompanied by a link to your code on github or bitbucket. Report should be in PDF (11pt; Times New Roman, Georgia or Arial). Submit your report via Canvas. Your report will be the source of your grade. It should also contain a link to your annotation file.

Report Template
You should use the American Medical Informatics Association (AMIA) template for your report. Word and LaTeX templates are available here (click on '+ Presentation Types'): https://www.amia.org/amia2019/call-for-participation.

Get Started
- Please email to request your annotation file. We will randomly allocate these files.
- Symptom dictionary: HERE

Grading
You will be graded on your ability to complete all 4 components. Approximately 40% of the grade will depend on the rule-based system and 60% on the other 3 components. Note that each component is important and the annotation component may impact your final system performance.

Due dates:

- Annotation submissions are due a week before the full submission is due. You will need to submit the annotated excel files.

- Full assignment submission is due at the end of week 5 (September 23). You will need to submit your report with link to your code on github/bitbucket.

  - System input:

- You will be provided with Reddit posts that will have a format almost identical to that of the annotations—only containing IDs, dates and texts. The system will have to load the file, process each post, and output the results.

- You will also be provided a separate set of Reddit posts that are already annotated, and you will report the performance of your system on that set (note: you cannot use the evaluation set in your system).

    - The inter-annotator agreement calculator script has been added to Canvas (IAA_Calculator.py)

- System output: an output format, similar to the annotation files, will be provided to you. We will try to make the output format as simple and easy as possible. Make sure your system's output matches the output format.

- System evaluation will only consider if a post (i) contains a specific symptom (*i.e.*, the evaluation script will look for CUIs or 'Other'/'C0000000'), and (ii) if the negation flag on the CUI. Please see Week 4, Lecture 2 (Thursday) and the IAA_Calculator.py to see how the negation flag can be appended to a CUI.

Weight: 15%

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, 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.