Tech Review

CS 410

Santiago Garcia Acosta (sg54)

An Overview of Text Analysis in Health Care

With the advancement of technology and the vast amount of medical literature detailing various experiments, results from medical cases, and general tests, we have come to a point where there is more medical data in the world than we seem to be able to actually use. One very interesting and potentially powerful implementation of this data that has become a subject of interest and effort over the past decade, is the idea of using text analysis methods to extract symptoms related to specific diseases and results from all of this medical data. This has an interesting application with disease classification. Imagine being able to use all of the past medical cases and experiments to check someone's symptoms and match them, with a high degree of accuracy, to some specific disease (or set of diseases) based off of a large amount of medical cases from the past few decades. We explore this topic, some of the work already done in the area, and what it may look like in the future throughout this review.

We can specifically look at a study done by Koleck et. al to better understand the applications of this concept in modern health science, through their paper "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review". Throughout the paper they specifically analyze a variety of other papers in the industry that have used NLP methods to extract symptoms from EHRs, electronic health records which detail real patients' information. They specifically found that the main use case for this application has been for disease *classification* tasks, rather than for understanding the symptoms themselves. This emphasizes the utility of these analysis tools for actual use in a clinical setting. One important thing to note, however, is that the paper was written in 2019, yet they only found 27 articles across PubMed that related to the topic of using NLP in some way with the notion of medical symptoms. This highlights that, while the technology exists and is accessible, its application and spread throughout the medical and bioinformatics realm has not really been strong.

With this in mind I want to focus on the possible applications of this technology, waiving away general legal and realistic semantics of course. The idea of integrating data science with the health realm is not a new one, yet the progress in that state has very clearly been pretty slow. This may be related to the very clear division of skills between the individuals that focus on the health field and those who focus on the use and understanding of such text analysis concepts, however it is hard to argue against the promise of such a merge. For example, in Jackson et al.'s paper, "Natural language processing to extract symptoms of severe mental illness from clinical text", they dive directly into such an integration. They specifically looked at previous EHR reports to develop a text analysis tool which could take a patient's symptoms and compute their closeness to specific mental health illnesses based on the previous EHR data. They found pretty convincing results and hoped that others would use similar steps for other symptomatic insights.

I believe that, with the number of patients doctors see every day (around 20), these tools can be used very effectively to train large models that can aid in the diagnosing process of medical visits. Notice that I say *aid* and not *replace*, as I strongly believe that such an application would work best in tandem with the doctors' expertise, enabling them both to help one another. Specifically, a doctor could submit a patient's EHR to the model, get back some ranked list of possible diseases that match those symptoms (all based off of other EHRs in the system) and the doctor could then use that information as a possible first guess. Once the final diagnosis is found, the doctor could mark the EHR with that diagnosis, and the model would use that label to make itself more intelligent. With the massive amount of medical cases across the world, this system could very easily and very quickly become incredibly strong in its statistical approaches, in a way that would not impede in the doctor's daily work.

In conclusion, we have seen that the use of text analysis techniques in medicine is not a novel concept, and is slowly growing. However, there is definitely much more that can be done with it, with a strong case existing for the large benefits it could provide for aiding medical personnel in diagnosing patients based on the symptoms that they present. One could argue that this could be extended with the patient's family medical history, or in a way that would enable the patient to get some answers before even going to the doctor. The main problem at hand is a lack of use, and therefore a lack of data, which could make saving people's lives and making people feel better, a much more expedited process.

References

1. Koleck, Theresa A et al. “Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review.” *Journal of the American Medical Informatics Association : JAMIA* vol. 26,4 (2019): 364-379. doi:10.1093/jamia/ocy173
2. Jackson RG, Patel R, Jayatilleke N, et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project