Natural Language Processing and the Data Problem in (Mental) Healthcare

Philip Resnik, University of Maryland

Session Summary

Philip Resnik, University of Maryland, gave the third keynote address of the 11th Proceedings of the Health Survey Research Methods Conference. He introduced himself as an outsider to health survey research, making the case for natural language processing in health care and health research.

Dr. Resnik opened his address focusing on the problem of mental health care. Direct and indirect costs from mental health care are higher than for cardiovascular disease, diabetes, cancer, chronic respiratory diseases combined. Yet, there are numerous challenges, ranging from recognizing people who are likely to attempt suicide, diagnosing conditions of psychosis, and even in accessing treatment. There is enormous potential for technology to help address these challenges. Dr. Resnik demonstrated this through a few of his studies, such as a Reddit suicide watch program aimed at identifying people who may attempt suicide. Despite the potential that machine learning can bring to mental health diagnostics and health care in general, natural language processing in health care is 10 years behind the state of the art. Our datasets tend to be small, and people are tentative in sharing data due to regulations such as HIPAA and concern for privacy and confidentiality of data. Progress in the field has been slow without shared data, because people are not able to work on the same problems.

Discussion

Naturally, the majority of the discussion following Dr. Resnik's presentation centered on how natural language processing can be incorporated into health survey research and the challenges in doing so.

A few participants posed questions around the resources necessary for carrying out these analyses. Specifically, questions centered on how much data or text is needed, whether the source and quality of the data reduce how much is needed, and what costs could be expected. Dr. Resnik's response was that more data is always better, but a lot of progress can be made even from relatively small amounts of text, such as tweets covering a period of a week or two. The Reddit suicide forum typically includes only one or two posts from a particular person, and their posts vary in length from just a sentence or two to long monologues. A follow-up question on this topic inquired as to whether having focus group data targeted to a specific topic can reduce the amount of needed data. Dr. Resnik responded that he has seen good correspondence between multiple focus group sessions and even a single focus group session.

The next part of the discussion focused on how we can make inferences to a population from social media data. Dr. Resnik's response is that we cannot just mine the data that is out there. People use the data they can get, but the future is not going to be in publicly available data. He went on to describe a project where they created a survey linked to other records and described a site (OurDataHelps.org) where people can choose to donate records for research. Surveys could include a final question that is an invitation to donate data. Someone else raised the point that documentation of the process will be important so that we know who ends up in the pool.

The last part of the discussion raised the issue of algorithm bias, in which the participant referenced a study that underestimated risks in minority populations. Dr. Resnik responded that this is a huge deal and an important interest. It was a wake-up call to the industry when Google photographs of black people started being labeled as gorillas. People in the industry are very aware of bias and are looking at ethics, including trying to use machine learning to detect bias. Yet, those in the AI community are not aware in the same way that social scientists are of areas such as sample bias. There is an annual conference focused specifically on these issues of bias and fairness (ACM (Association for Computing Machinery) Conference on Fairness, Accountability, and Transparency (ACM FAccT), https://facctconference.org/).