Research Seminar: How to Make Causal Inferences Using Texts
Speaker: Brandon Stewart, Princeton University, Sociology
Day and Date of Seminar: Wednesday, November 29th, 2017

In this seminar, the speaker talked about how texts are increasingly used to make causal inferences: either with the document serving as the treatment or the outcome. He reiterates the basic concept behind text data and causal inference, their existence usually in natural combination and how text analysis poses many problems for causal inference because the data are high-dimensional – text can express many different concepts and can express the same concept in many different ways.

Speaker's research suggests a tradeoff between discovery and testing in causal inference with text as treatment and outcome and provides a simple solution. Further explaining why developing the mapping between the text and categories of interest in the same set of data that is used for measuring a treatment effect violates the Stable Unit Treatment Value Assumption (SUTVA) commonly invoked in causal inference through a concept called as "analyst-induced SUTVA violation" (AISV). Speaker's research suggests the solution to this problem with a simple split of the data that allows the researcher to freely define measures of interest through an exploratory analysis of one dataset. It also suggests testing these measures of interest on the held-out data to confirm that the relationships are not an artifact of exploration. This train-test split is similar to machine learning and A/B tests in industry, which approximates a sequential notion of science where first a finding is discovered and in a separate experiment confirmed. In their (Speaker and Co-researchers) research, they apply this process of sequential science to two applications, one where text is the outcome of a survey experiment, and one where text is the treatment affecting a response.

The speaker clarified the issue with the analysis of high-dimensional data like text that requires the researcher to discover the mapping between the data and the quantities of interest which ultimately undermines the researcher's ability to make causal inferences. In his research, he clarified this issue by describing the Analyst-induced SUTVA violation and show that this type of SUTVA violation can be solved with a simple split of the dataset into a training set used for discovery and a test set used for testing. Also, the analyst-induced SUTVA assumption does not only apply to text data, but to any latent representation of treatment and outcome variables. The research also advocates the use of train/test split framework in the context of other types of data that require a low-dimensional representation of high-dimensional data.

One of the many parts of the seminar that stood out to me was about stepping back from assigning any truth-value to one given study. The root idea behind a research is to discover new findings or re-evaluate the result of preexisting ones, but some/many are still being operated under the mindset of producing the absolute truth, "infallible results". If given result can't be taken as true or real on its own, we should work on the research as if we believe that and work

further to leave space for any further advances. Practicing with that belief in that collective model of producing knowledge, on an individual level, is structuring research in a modular way and doing things like train-test splits. The logical outcome of a good methodology is that even when wrong, your methods can adapt themselves to eventually be right. The logical outcome of this kind of adaption could be that research can never wrong, only steps away from being right.

Text analysis is one the methodologies, that I would employ to explore a topic in a field that I aim to study further. While learning about the methods to capture, analyze and provide a cumulative solution to the growing problem of cyberbullying in the social media generation, the knowledge about employing various techniques to analyze the high-dimensional, complicated, and sparse data like text, can drastically improve the results and create a space for researchers to try out different models and methods.

- Tapasvi Bansal