Text as Data (Gentzkow, Kelly, and Taddy 2019)

Computers can identify emails as *spam* or predict whether a comment is negative or positive. Both aim to predict an outcome – if an email is spam or not for the first example – using text as data.

One Sentence Summary

This paper explains four methods for using text as data for economic research. It also shows how could be manipulated for causal analysis and with the statistical and computing principles behind each method.

Main Findings

Economists aren't used to consider text as data, mainly because of its size. If we wanted to use a twitter post with all possible combinations of words, the data could potentially have the same size as the number of atoms in the universe. On the last decades statisticians have tried their best to do causal inferences from few data points, and little has been done on statistical analysis for monstrous data sets.

Now, computers can do the heavy lifting. First, we need to reduce the *dimensionality* – or size – of text in the data. Then we can apply a statistical method for inference – that will depend on the research question.

Each language has several connectors and repeated words that embellish a message. However, it is possible to boil down a text to a set of significant words. We can reduce dimensionality if we strip away unnecessary words. For this we can: i) use "stop-words" approach to remove words like "the" and "a" or "and", ii) exclude both common and extremely rare words, and/or, iii) use "stemming" to simplify complex word like economically to economic.

Once we have a "manageable" – but still big – data set we can use a method for causal inference. Consider the next question: How much do soft skills matter in a sales pitch? The ability to persuade another to buy a certain product, could be related to non-technical knowledge. Meaning that sales teams has soft skill, which improves their chances of a successful sale. We can use the sales pitches to see how soft skills influence sales success (Macarena, 2019).

A *dictionary* approach is the easiest method for text economic analysis. From theory, a variable of interest could be constructed by counting or grouping sets of words. For example, a theory argues that words like "poor" or "sad" could be used to influence a potential buyer. Then the index could be the number of words that influence the potential buyer in each sales pitch. As a final step, we can use this data for typical casual analysis in economics.

If an individual has a successful sale, then he/she can learn and increase their soft skills. A *generative models* can use the probability of success to determine how their speech will change.

This approach fits well when predicting how a speech could adapt to new audiences or new buyer demographics.

Concluding Remarks

Computing power has opened the gates for new and interesting statistical methods for economists. Most are skeptical of them, with good reason. However, they expand the amount of questions that can be answered, such as: what are the key words in a speech that increases candidate possibility of being elected? Which can lead us to understand an unexplored side of voting behavior. Text analysis can open unexplored knowledge of current puzzles in economic thinking.

References:

- Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. “Text as Data.” *Journal of Economic Literature* 57(3):535–74. https://www.aeaweb.org/articles?id=10.1257/jel.20181020.
- Macarena, Rosario. 2019. “The Power of Asking Right: A Field Experiment.” Working Paper. SECHI, Universidad de los Andes, Santiago, Chile on 8/13/2019.