Report: Replicating “Measuring and Explaining Political Sophistication through Textual Complexity”

PPOL 6801, Spring 2024
Prof. Tiago Ventura
Zhiqiang Ji

Summary

This replication study reproduced (Benoit et al., 2019). The original paper, "Measuring and Explaining Political Sophistication through Textual Complexity," addresses a challenge faced by political science researchers: despite the critical role of political communication in political science and a longstanding interest in complex political communication such as dog whistles and political rhetoric, there has been a lack of suitable tools to measure the textual complexity that captures the unique sophistication of political texts. The original study utilized crowdsourcing to conduct thousands of pairwise comparisons of text snippets from State of the Union addresses, integrating these assessments into a statistical model of sophistication. Since the manual coding work of crowdsourcing is not replicable, this replication study used the already coded text data to fit the model. The mathematical simplicity of the Bradley-Terry model fitting and the significant impact of key covariates such as word rarity resulted in consistent outcomes when selecting important variables using random forest. Consequently, this replication essentially reproduced the original study's findings.

Process

The original authors generously provided all the necessary data, well-annotated code, and documentation, which offered crucial guidance for the replication effort. However, this replication study encountered difficulties that the original authors might not have anticipated.

Firstly, in terms of software environment setup, in the years since the publication of the paper, some software packages had changed due to updates, others were no longer actively maintained, and some could not be easily installed on new hardware platforms, such as Apple's silicon ARM64 architecture CPUs.

After setting up the software environment, another technical challenge of the replication study was organizing the sequence in which different parts were replicated. This was because some steps required significantly more computational time than others. The most time-consuming task in this replication—bootstrapping four fitted structured Bradley-Terry models—took about 30 hours. (Thankfully, the authors indicated in the project documentation which parts were time-consuming and the approximate duration needed.) The longer a task runs on a personal computer, the higher the risk of an unexpected interruption. It's best to ensure no low-level errors like typos in the code before running it.

Moreover, while waiting for such lengthy tasks to be completed, writing and revising other parts of the R notebook without performing calculations or picking up other tasks that would facilitate the rest of the replication became necessary alternatives. For instance, since the starting point of this study was data manually labeled through crowdsourcing—a process I couldn't replicate—I could use the waiting time to understand how the authors prepared the data, managed the quality of manual labeling, and integrated the final dataset generated by the crowdsourcing work. This replication was my first encounter with waiting for long tasks to run on my local computer. It was a valuable learning opportunity, allowing me to think about multitasking effectively during replication. Another insight from the time-consuming tasks was the importance of saving intermediary results that take a long time to generate at each step, effectively avoiding the trap of repeatedly rerunning a lengthy task.

The rest of the replication work, thanks to the comprehensive and detailed guidance provided by the authors, did not present significant difficulties. However, the thorough instructions left little room for new methods or experimentation. Especially since I consider myself a novice in statistics and NLP, the replication was more about learning and understanding the theories and software tools used by the authors. Perhaps unlike many replication studies, my main challenge was not the difficulty in reproducing the original results—under the authors' scaffolding, it seemed unlikely to arrive at drastically different outcomes. However, my main challenge was to continuously confirm that I truly understood how the results were obtained or why they needed to be obtained in a specific way.

Differences

There were minor discrepancies between the original work's results and my own, but nothing was surprising or implausible. All of them could be due to the selection of random seeds and slight differences in the parameters during model fitting.
However, I discovered a possible typo in the original paper, which I also noted in the code notebook. Table 2 of the original paper seems to have the `Prop correctly predicted` values for `FRE Reweight` and `Basic RF Model` reversed. Although these two values are very close (0.737 vs. 0.738), the annotations in the original code and the results of my replication both indicate that `FRE Reweight` is slightly higher.

Autopsy

In this section, I would like to discuss the lessons I learned. I noticed gaps between the theoretical framework presented in the paper and its practical implementation in the code. Although the idea behind the new model is very straightforward—first estimating the easiness of each article using an Unstructured Bradley-Terry model, then adding many intuitive predictors that are either widely used or newly introduced, and finally using a random forest to select the most important predictors. When I first read the paper, the simplicity of the approach made me wonder why such a measurement hadn't been developed in the past decades.
However, understanding this approach does not mean that I am crystal clear on the theoretical choices made by the authors and why they ultimately chose the current scheme, such as why they returned to the B-T model after finding the key predictors to complete the final model framework, or why they did not fully comply with the RF's pick and chose the top 3 predictors plus the 9th important one to form the best model. Secondly, being able to get the outputs does not mean I fully understand the technical side of data processing. For example, if I were replicating the same research with less guidance, I wouldn't compare the cases with and without bias reduction when fitting the unstructured B-T model for the first time. Because the code and documentation provided by the authors were very detailed, it was easy to replicate similar results, which made me feel that I might "not know what I don't know" due to a lack of relevant knowledge. This replication made me deeply realize that a seemingly simple and intuitive idea

requires many behind-the-scenes decisions to be implemented thoroughly and robustly, and these decisions are often left in the code. This also shows that replication provides a more hands-on learning opportunity than ordinary reading.

Extension

First, I believe AI has the potential to significantly aid in the preliminary stages of research by partially substituting human judgment. Recent advancements in technologies like Retrieval-Augmented Generation (RAG) by OpenAI and other large language models (LLMs) enable AI to repeatedly retrieve information from a corpus and perform text analysis without the need for continuous input of text, akin to sending queries. This approach could be exponentially faster, cost-effective, and potentially more consistent quality when coding with "semi-human judgment" across a large corpus than crowdsourcing companies.

Second, the author's suggestion to make all variables dynamic, such as measuring the proportion of nouns in a text relative to a baseline from the time the document was written, opens up exciting avenues for future research. Extending their work to include combinations of grammatical and syntactical features could allow for measuring metrics beyond readability. The flexibility of this framework makes many types of measurements possible. For instance, increasing the weight of noun usage could help analyze the complexity of congressional bills, or examining the proportion of verbs might correlate with the persuasive power of a candidate's speeches. Naturally, using corpora from different domains—such as news reports, novels, and social media conversations—this framework that combines various linguistic elements with the B-T model could have a wide range of applications. By leveraging the framework's flexibility and incorporating dynamic variables, researchers could develop more nuanced and historically contextualized measures of text complexity or other linguistic features and provide a richer understanding of communication and expression in different settings.