Scaling the Meaning of Text with Word Embeddings

Proposal for the Cambridge Elements Series on Quantitative and Computational Methods for the Social Sciences

June 2024

Authors

Colin R. Case
Assistant Professor, Department of Political Science, University of Iowa
colin-case@uiowa.edu

Rachel Porter
Assistant Professor, Department of Political Science, University of Notre Dame
rachel.porter@nd.edu

Synopsis

Political phenomena are commonly thought of in spatial terms. To place political actors in space, researchers use actors' observed behavior(s) to scale their positions along dimension(s) of interest relative to other political actors. This line of study often involves measuring positioning on a unidimensional axis with poles indicative of a substantive quantity of interest (e.g., support versus opposition, positivity versus negativity, civility versus incivility, or liberal versus conservative) where an actor's proximity to a pole indicates her closeness to that position or preference.

Digitization has broadened the accessibility of data suitable for scaling political actors' positions and preferences. Text, in particular, has proven to be a promising source of data as a variety of latent concepts are embedded in and can be extracted from text. Machine learning methods for scaling positions from text, such as wordfish (Slapin and Proksch, 2008) and wordscore (Laver et al., 2003), have seen widespread use for this reason. These tools, though, have some notable limitations. Bag-of-words approaches for scaling text neither account for the contextual use of words nor the semantic relationship between words. Moreover, Egerod and Klemmensen (2020) show that these approaches often violate assumptions foundational to scaling text (e.g., conditional word independence). Grimmer and Stewart (2013) argue that unsupervised bag-of-words scaling techniques can produce unstable estimates of latent constructs if they incorporate no external information in estimation.

Recent advances in word embedding techniques for scaling text offer a promising path forward. Word embedding models employ a neural network architecture to predict word(s) in a document given the word(s) that occur in close context to that word. Model output embeddings are vector representations of semantic relationships between words in a dense, continuous, high-dimensional space. Words with similar contextualized use are located proximate to one another in space, and words with more dissimilar use are located farther apart. Additionally, relationships between embeddings can be expressed using algebraic operations (Collobert and Weston, 2008). For the purpose of scaling, this means that we can calculate the orthogonal projection of some latent dimension by subtracting word embeddings representing the poles for that quantity of interest (e.g., $\overrightarrow{\text{support}} - \overrightarrow{\text{oppose}}$). The result of this computation is an "axis" of meaning, and text can be scaled along this dimension by calculating the cosine similarity between a given word embedding and the axis embedding. This approach for semantic projection, first introduced in work by Bolukbasi et al. (2016), has been applied to questions spanning multiple social scientific fields—including sociology, psychology, and linguistics—with far fewer applications seen in political science. This may be because extant work has focused on capturing word-level latent dimensions, such as cultural associations (Kozlowski et al., 2019), stereotyping (Garg et al., 2018), and physical properties (Grand et al., 2022). Conversely, research in political science is more often interested in capturing latent dimensions at a higher level of aggregation (e.g., documents or covariates).

This book provides an accessible guide to scaling text using a word embedding approach. We begin by describing a motivating problem within the field of political science. Specifically, we review the challenges associated with scaling text for politicians' left-right positioning on policy (e.g., taking a pro-life versus pro-choice stance on abortion); this serves as our running example throughout the book. We go on to discuss best practices for selecting corpora, extracting text of interest from documents using machine learning classifiers, and estimating word embedding models. Next, we discuss the logic behind semantic projection and explore approaches for producing an axis of meaning. In doing so, we put forth a framework for scaling the meaning of text at the document or covariate level. Subsequent chapters explore some potential avenues for evaluating measurement robustness and validation. In particular, we explore the utility of open-source and proprietary LLMs as a validation method. We also benchmark the performance of semantic projection relative to bag-of-words approaches against gold-standard human evaluations. Finally, we discuss extensions and other applications for this text scaling approach.

Key Contributions

We view the contributions of this book as fourfold. First, we provide scholars with the intuition and tools to implement embedding-based text scaling in their work. Word embedding techniques provide many exciting opportunities for research; however, scholarship spans multiple sub-disciplines, making it challenging for newcomers to become acquainted with the work. We provide a survey introduction to embedding-based approaches for scaling text and seek to sufficiently familiarize readers so that they can go on to engage with frontier scholarship in this fast-moving space. Second, we provide readers with an extended, detailed example of a machine-learning pipeline for text scaling at the document and covariate level. Third, we underscore the critical importance of checking robustness and validating text-based measures. As Park and Montgomery (2023) document, there is a concerning lack of standardization and validation in current text-to-measurement research. As Rodriguez and Spirling (2022) note, outputs from word embedding models are more robust but still sensitive to researcher degrees of freedom and model specification choices. We hope our carefully executed example of a text-to-measure pipeline can inspire readers' work. Finally, the measures we produce for congressional politicians' policy-specific positions are themselves novel. Extant measures that summarize politicians' positioning using a single ideology or positioning "score" are not well-suited to answer many pressing questions about issue-based polarization, electoral choice, and policy representation. Our measures address this important gap.

Proposed Book Plan

Introduction

Key Concepts

- Broad discussion of traditional and text-based spatial approaches for scaling
- Challenges associated with bag-of-words approaches for scaling text
- Introduction to word embeddings and the logic of semantic projection
- Structure of the book

In this chapter, we introduce spatial approaches for scaling, beginning with a review of traditional approaches (e.g., hand-labeled data, IRT models). We then go on to discuss popular text-based approaches (e.g., wordfish, wordscore) and their limitations. Finally, we provide a broad overview of word embeddings and introduce the intuition underlying semantic projection.

Chapter 1 - Estimating Embedding Models

Key Concepts

- Provide a brief overview of word embedding model parameters and tuning choices
- Engage in a more in-depth discussion of approaches to representing documents
- Provide a guide for splicing text into the desired unit of aggregation

This chapter focuses on the choices one must make when estimating a word embedding model. We first engage with specific parameter and tuning choices that must be made before model fitting that are universal to embedding approaches (e.g., window size selection, embedding dimension specification, pre-trained versus locally trained). Extant literature has explored the properties and performance of models fit with various specifications (e.g., Rodriguez and Spirling 2022), so we keep our summary brief. We go on to more deeply discuss multiple approaches for representing documents or covariates at the embedding level (e.g., producing embeddings for speeches in a corpus or candidates in an election). In this discussion, we use our running example to demonstrate these different aggregation approaches (e.g., averaging all embeddings in a document versus employing a doc2vec architecture). This section also walks through the process of disaggregating texts into smaller documents using a machine learning classifier (e.g., for our purposes, splicing a complete campaign platform into policy-specific documents).

Chapter 2 - Defining the Axis of Meaning

Key Concepts

- Describe the logic underlying semantic projection
- Discuss strategies for selecting anchor words to define poles of interest
- Outline process of estimation

This chapter explores the principles underlying semantic projection and provides a walk-through for producing an axis of meaning. We underscore the importance of unbiased selection for anchor words (e.g., words indicating support versus opposition) and highlight algorithms that achieve this aim (e.g., King et al. 2017; Schopf et al. 2022). Using our running example, we outline the pipeline for producing an axis of meaning.

Chapter 3 - Robustness & Validation

Key Concepts

- Walk through the procedure for scaling text onto axis of meaning
- Internal and external validity checks
- Discussion of trade-offs associated with various approaches to representing documents

In this chapter, we engage in the final step for scaling text using semantic projection. After doing so, we use our running example to provide a template for validating our text-based measure's external and internal validity. In our specific application, we use the crowd-sourced pairwise comparison framework proposed by Carlson and Montgomery (2017) to benchmark our measure against human judgments. We additionally compare our policy-specific positioning scores to fundraising differentials from single-issue PACs (e.g., do extreme pro-life candidates receive more contributions from the National Right to Life Committee?). We also assess the relative performance of those various approaches to representing documents discussed in Chapter 1. This includes a discussion of the trade-offs associated with each approach and their most appropriate use cases.

Chapter 4 - Benchmarking Against Alternative Approaches

Key Concepts

- Compare semantic projection to wordscores, wordfish
- Compare semantic projection to LLM-based measurement

This chapter demonstrates the advantages of semantic projection; we replicate our estimation procedure using alternative approaches for text scaling. Using our running example, we benchmark these measures against our own and the human judgments discussed in Chapter 3.

Chapter 5 - Extensions

Key Concepts

- Sentiment approach - economic sentiment
- Unipolar approach - bipartisanship
- Multidimensional approach - government

In this chapter, we demonstrate extensions to our approach for semantic projection. In our first example, we use well-established dictionaries of meaning to scale sentiment towards the economy (i.e., positive versus negative); our goal is to demonstrate that semantic projection is not specific to policy-oriented questions. In our second example, we use a unipolar axis of meaning to measure the use of bipartisan language in documents; our goal is to demonstrate that semantic projection can also uncover rhetorical style. In our third example, we demonstrate the use of this method to scale words across multiple dimensions simultaneously to create multidimensional scores.

Chapter 6 - Conclusion

The concluding chapter reflects on the fast-moving state of text-to-measurement approaches in the social sciences. We discuss potential areas for future research and wrap up considerations for researchers. Finally, we discuss coding and machine learning pipeline resources that we make available for a general audience of researchers.

Audience

This book will be of interest to a broad audience of social scientists, including political scientists, who are interested in methods for text-to-measurement. Our book will be accessible to audiences with and without familiarity with this subject matter. To increase the accessibility of our book to novice readers, we intend to embed code throughout the book itself and provide replication materials accessible through Code Ocean.

Timeline

Our draft book manuscript will be prepared for submission to peer-review by December, 2024.

References

Bolukbasi, T., K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems 29*.

Carlson, D. and J. M. Montgomery (2017). A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review 111*(17), 835–843.

Collobert, R. and J. Weston (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167.

Egerod, B. C. and R. Klemmensen (2020). Scaling political positions from text: Assumptions, methods and pitfalls. In *The SAGE Handbook of Research Methods in Political Science and International Relations*, pp. 498–521.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences 115*(16), E3635–E3644.

Grand, G., I. A. Blank, F. Pereira, and E. Fedorenko (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behavior 6*, 975–987.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*(3), 267–297.

King, G., P. Lam, and M. E. Roberts (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science 61*(4), 971–988.

Kozlowski, A. C., M. Taddy, and J. A. Evans (2019). The geometry of culture: Analyzing meaning through word embeddings. *American Sociological Review 84*(5), 905–949.

Laver, M., K. Benoit, and J. Garry (2003). Extracting policy positions from political texts using words as data. *American Political Science Review 97*(2), 311–331.

Park, J. Y. and J. M. Montgomery (2023). From text to measure: Creating trustworthy measures using supervised machine learning.

Rodriguez, P. L. and A. Spirling (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *Journal of Politics 84*(1), 101–115.

Schopf, T., S. Klimek, and F. Matthes (2022). Patternrank: Leveraging pretrained language models and part of speech for unsupervised keyphrase extraction. *arXiv preprint arXiv:2210.05245*.

Slapin, J. B. and S.-O. Proksch (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science 52*(3), 705–722.