Mining Meaning: Semantic Similarity and the Analysis of Political Text

Marika Landau-Wells*

Working Paper, January 15, 2023

Abstract

The degree of similarity in meaning between texts (e.g., manifesto items, speeches) is often of fundamental interest to political scientists. Categorizing texts based on meaning, instead of dictionary-based matching, requires solving the qualitative problem of "what goes with what." In this note, I show how a pre-trained language model optimized for semantic textual similarity can help provide independent validation for researchers solving this problem. I introduce a new measure of discriminability – relative semantic similarity (RSS) – that captures how coherent any category of texts is in terms of its semantic meaning, relative to another category. Using the pre-trained model's output, I show that RSS can be used as a test statistic to (1) independently validate the coding scheme of a manually categorized corpus, and (2) test for confounders that might affect the distribution of semantic meaning within a corpus. RSS thus complements and extends the text analysis toolkit for social science.

Word count: 2985 (including references)

*Travers Department of Political Science, University of California, Berkeley (mlw@berkeley.edu). I would like to thank Emily Gade, Aidan Milliff, Rich Nielsen, and Eric Schickler for valuable comments.

The degree of similarity in meaning between texts (e.g., official statements, manifesto items, speeches, free response items) is often of fundamental interest to political scientists. To categorize texts along the dimensions they care about, scholars create code-books and other schemes to consistently define "what goes with what" (e.g., The Comparative Manifesto Project's Coding Handbook) (Werner et al. 2021). In principle, if other human coders follow these same guidelines, then they should generate highly similar categorical judgments (Mikhaylov, Laver, and Benoit 2012). Yet, manual coding is also vulnerable to confirmation bias (Chakrabarti and Frye 2017), variability in coder expertise (Klingemann et al. 2007), and task difficulty (Mikhaylov, Laver, and Benoit 2012).

In this note, I introduce a new text analysis measure, *relative semantic similarity* (RSS), designed to aid researchers faced with categorizing "what goes with what." RSS helps researchers quantify the extent to which a categorization scheme picks up on independently discriminable semantic nuance within a corpus. RSS is not an alternative to human coding or qualitative assessment. Nor is it a text classifier. Rather, RSS is a measure that can be reported to signal the robustness of a manual coding scheme that relies on distinctions in *semantic meaning*. Semantic meaning refers to meaning that is contingent on both syntactic construction and lexical choices (Lappin 2017). RSS can also be used to test hypotheses about the distribution of semantic meaning in a corpus.

Since Grimmer and Stewart (2013) initially laid out the benefits of using text as data, scholars have embraced and extended methods developed in computer science and adjacent fields to better understand and analyze political texts (for reviews, see Wilkerson and Casas 2017; Benoit 2020). Chatsiou and Mikhaylov (2020) note that natural language processing (NLP) models, which combine computational linguistics and deep learning, are particularly promising for political science.

RSS uses one such NLP model – a freely available, pre-trained language model optimized for the task of *semantic textual similarity* (STS) (Reimers and Gurevych 2019). This type of NLP

model requires no additional training data or technical expertise to use. It operates on text *strings*, not just words, and encodes those strings into fixed-length numeric representations in a process known as vectorization or embedding.[1] These embeddings capture syntactic, semantic, and entity information (Rogers, Kovaleva, and Rumshisky 2021). With STS-trained models, the closer two texts' embeddings are in the model's representational space, the more semantically similar they are likely to be.[2]

Are these models good enough to be useful? I show several benchmarking analyses in the Online Appendix, but the main take-away is that the model I use achieves a correlation of 0.91 with human raters' similarity judgments; it can detect negation; and it ignores the superficial similarities that stymie word-level encoders like Word2Vec (Mikolov et al. 2013). For reference, Ruedin and Morales (2019) reported a correlation of 0.86 between expert survey respondents and manual coders on a one-dimensional judgment task. And Benoit et al. (2016) report a correlation of approximately 0.95 between expert raters and crowd-sourced non-experts on a three-category classification task.

As Rodriguez, Spirling, and Stewart (2021) note, a major challenge with using encoding models for inference is that their output is not directly interpretable without "some notion of a null hypothesis, some understanding of the variance of our estimates, and a test statistic" (3).

The measure I introduce, *relative semantic similarity*, fulfills these criteria. The intuition for RSS is that the semantic meaning of a text *should be* more similar to those texts *within* the same category than to those in other categories. RSS captures how true this statement is for any two categories of text. In the next section, I show that RSS has a directly interpretable

[1] I discuss encoding models in greater detail in the Online Appendix. The STS-trained model I use is `stsb-mpnet-base-v2` (Reimers and Gurevych 2019), available at https://huggingface.co/sentence-transformers.

[2] Proximity is calculated as the cosine between the two vectors ("cosine similarity"). Cosine similarity is the standard measure of similarity in NLP tasks (Reimers, Beyer, and Gurevych 2016) and usually ranges from 0 to 1 in the case of text (Benoit 2020).

null value and can be used as a test statistic within a permutation inference framework.

I then provide two use-cases for RSS. In the first, I demonstrate how RSS can be used to *independently* validate a qualitative coding scheme using an original corpus of Cold War documents hand-coded for subtle linguistic distinctions. In the second case, I show how RSS can be used to test for confounders that might affect the distribution of semantic meaning within a corpus. In both cases, the similarity judgments provided by the model are entirely replicable, unlike those provided by human coders (Mikhaylov, Laver, and Benoit 2012), and are unaffected by the researcher's own biases or priors, as long as the model is applied without adjustment.

All methods discussed in this paper can be implemented on a laptop in Python or in R with a Python installation. A minimal working example in R is available via the Online Appendix.

Relative Semantic Similarity

How similar is one text to any other? When categorizing documents, paragraphs, sentences, free responses or tweets, researchers often rely on their expertise and intuition to render subtle linguistic judgments. But underlying all such judgments is a shared claim: all texts in one category are defined as being fundamentally *more similar to one another* than they are to texts in another category on the dimensions relevant to the categorization scheme.

A mutually exclusive categorization scheme relies on the assertion that the texts in Category **X** are more semantically similar to one another than they are to texts in Category **Y**. This property of categorical distinctiveness – that *within*-category similarity should be greater than *between*-category similarity – was first illustrated for high-dimensional (neuroimaging) data by Haxby et al. (2001). In the case of text, STS-trained encoding models can produce the necessary similarity judgments. Specifically, for any categorized collection of sentences

encoded using the model, the semantic similarity of each sentence to all others can be calculated from the inner product of their two embedding vectors (i.e., their cosine similarity).

It is relatively rare for text to be annotated at the sentence level, however. Researchers may wish to consider longer passages or entire documents. Fortunately, vector representations of sentences can be averaged into vector representations of longer spans of text (Bojanowski et al. 2017). Thus, any text can be represented within the model's feature space, either by encoding it directly, or averaging the embeddings of its components.

I combine the representation of texts in a shared similarity space with the principle of categorical distinctiveness to generate a new measure: *relative semantic similarity*. The full derivation appears in the Online Appendix. But, in summary, the RSS for Category **X** with respect to Category **Y** is defined as the difference score for the average *within*-category similarity for Category **X** (W_X) and the average *between*-category similarity for Categories **X** and **Y** (B_{XY}):

$$WB_{XY} = W_X - B_{XY} \tag{1}$$

If the value WB_{XY} is positive, then Category **X** texts are discriminable on average from Category **Y** texts. That is, they are more semantically similar to each other, on average, than they are to texts in Category **Y**. The opposite is not necessarily true, as it relies on the relative coherence of Category **Y**, i.e. W_Y. Where two categories are not semantically discriminable, the difference score WB_{XY} is not distinguishable from zero. Where one category is particularly incoherent, from a semantic perspective, the value of WB_{XY} could be negative.

It is important to note that while the range of cosine similarity values calculated for a corpus is a function of the chosen encoding model, and thus, somewhat arbitrary, the WB_{XY} measure is directly interpretable. Significance testing can be performed using WB_{XY} because

the null hypothesis of no difference in semantic meaning between categories is represented by $WB_{XY} = 0$. The null distribution of any two (or more) categories can be derived by permutation, i.e., shuffling the labels of Categories **X** and **Y** and recalculating WB_{XY} each time. The observed WB_{XY} value can then be compared to the null distribution where the likelihood of a false positive will be defined by the number of permutations (Ernst 2004). Thus, WB_{XY} can be used as a test statistic, and it allows researchers to make inferences of the kind Rodriguez, Spirling, and Stewart (2021) advocate, e.g., testing categorical differences across subcorpora.

Use-Case 1: Validating Qualitative Coding Schemes

I use a new dataset of documents from the early Cold War concerning Communism (the Countering Communism Corpus) to illustrate how RSS can be used to validate a hand-coding scheme. In brief, the CC Corpus contains texts in which American policy-makers discussed the threats posed by Communism. It spans the period 1939-1953 and contains 289 documents by 22 authors. These 289 documents contain 12,263 paragraphs and 38,564 sentences. Each paragraph in the corpus has been hand-coded for a qualitative assessment: does the paragraph contain discussion of Communism as an existential threat (Category 1), as a threat to rights and institutions (Category 2), as a virus-like, dangerous idea (Category 3) or is there no discussion of Communism as any kind of threat (Other text, coded 0)? I provide examples of each category in the Online Appendix.

I encoded the 38,564 sentences using an STS-trained model (`stsb-mpnet-base-v2`). I then generated 12,263 paragraph-level embeddings by feature-wise averaging. Panel A of Figure 1 shows the average semantic similarity between all threat-related paragraphs (Category 1, 2, or 3) and all Other paragraphs (coded 0) in a symmetric matrix. On average, all paragraphs in which the danger of Communism is discussed are more similar to one another than to paragraphs discussing other topics. Panel B of Figure 1 breaks apart the threat-related

paragraphs to validate the semantic distinctiveness of the three-category hand-coded scheme. Within-category similarities are shown on the diagonal, and between-category similarities are the off-diagonals. Panel C shows the results of significance testing using RSS across all category pairings. As Panel C shows, all observed RSS values from Panel B (vertical lines) lie above the null distribution with 500 permutations, which suggests that the qualitatively-defined hand-coding scheme captures differences in semantic meaning that can be recognized by an independent coder and are unlikely to be false positives ($p < 0.002$ in all cases). While the differences in Panel B are relatively small on the scale offered by cosine similarity, Panels A and C put the model's achievement in perspective. Within an already-coherent macro-category (threat-related content), the model also confirms there are additional subtleties of meaning.

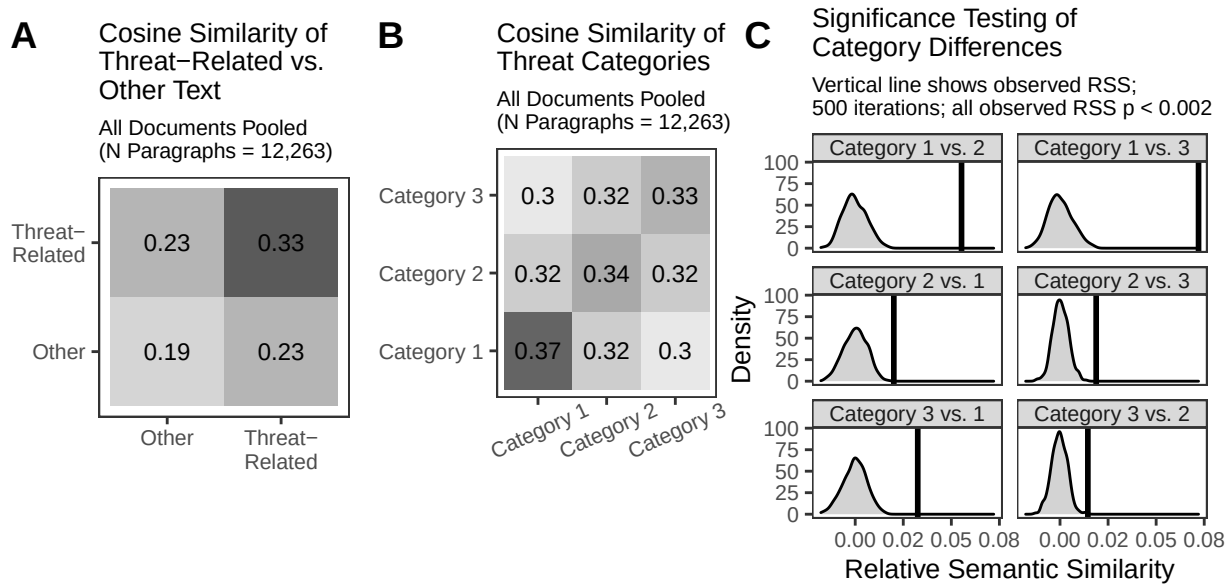

Figure 1: Coding Category Validation

Use-Case 2: Testing for Confounders

A logical extension of testing for *desirable* differences between subsets of text (e.g., coding scheme validation) is testing for *undesirable* differences. Undesirable differences might arise due to the presence of a confounder. Some confounders affect the frequency with which texts

are produced (Roberts, Stewart, and Nielsen 2020). Others (e.g., a document's intended audience) might affect semantic content directly in ways that jeopardize general claims about a corpus.

In the case of the CC Corpus, one might think that the secrecy status of a document has an effect on how individuals express their assessment of Communism as a threat. Text intended for wide public consumption (Never Classified) might thus differ systematically in content and meaning from material designated Top Secret or Off-the-Record. Such a confounder would invalidate unconditional inferences about the use of threat-related language in the corpus.

Using document-level embeddings, also generated by featurewise averaging, and the same permutation-based approach, I test whether or not secrecy status matters for the semantic content of threat-related text. Figure 2 shows the results of permuting the three secrecy status labels on the threat content of all documents and calculating the six WB_{XY} values each time. For each WB_{XY}, the observed value (vertical line) lies within the null distribution such that it is unlikely that WB_{XY} is meaninfully different from zero. See the Online Appendix for exact p-values. Based on this analysis, the threat content of Top Secret documents is not semantically distinguishable from the threat content in Never Classified or Off-the-Record documents. Thus, it is not necessary to condition claims about the threat-related content of the CC Corpus on secrecy status. Moreover, we can reject the hypothesis that classified documents contain qualitatively different threat language than unclassified documents in this corpus.

Conclusion

Scholars often go through long and laborious processes to justify their qualitative categorization of texts (e.g., Klingemann et al. 2007). In this note, I have introduced *relative semantic similarity*, a new measure that takes advantage of advances in NLP to complement or replace

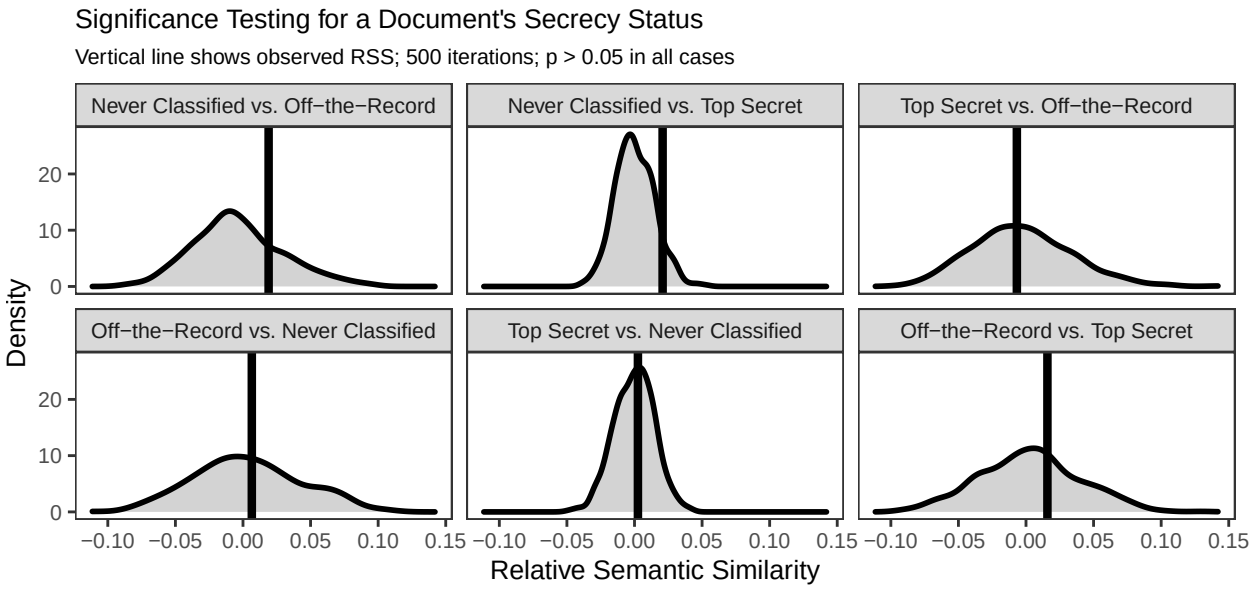

Figure 2: Testing for Confounders

the multi-coder model of validation. RSS provides scholars with an interpretable quantity of interest when combined with the outputs of STS-trained sentence encoders. I demonstrated how RSS can function as a test statistic to assess whether a qualitatively defined coding scheme is picking up on an independently observable differences in semantic meaning. I also showed that RSS can be used to enhance confidence in claims that the distribution of semantic meaning within a corpus is not influenced by a confounder. Both of these use-cases have value if we are concerned that researcher biases or coding scheme complexity affects the hand-coding of text. The model's similarity judgments are also perfectly replicable.

RSS can also be extended to other types of similarity relationships captured by specialized encoders. While the focus in this note was on similarity between short spans of text, there are other specialized encoders that match questions to answers or topics to posts. RSS provides a method for assessing human judgments (or deriving most-likely pairings) for any of these matching tasks. In sum, RSS provides a quantitative complement to human judgments about "what goes with what." These judgments are often central to our understanding of political texts.

References

Benoit, Kenneth. 2020. "Text as Data: An Overview." In *The SAGE Handbook of Research Methods in Political Science and International Relations*, edited by Luigi Curini and Robert J. Franzese, 461–97. SAGE Publications Ltd. https://sk.sagepub.com/reference/the-sage-handbook-of-research-methods-in-political-science-and-ir/i4365.xml.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–95. https://www.cambridge.org/core/journals/american-political-science-review/article/crowdsourced-text-analysis-reproducible-and-agile-production-of-political-data/EC674A9384A19CFA357BC2B525461AC3.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." http://arxiv.org/abs/1607.04606.

Chakrabarti, Parijat, and Margaret Frye. 2017. "A Mixed-Methods Framework for Analyzing Text Data: Integrating Computational Techniques with Qualitative Methods in Demography." *Demographic Research* 37: 1351–82. https://www.jstor.org/stable/26332229.

Chatsiou, Kakia, and Slava Jankin Mikhaylov. 2020. "Deep Learning for Political Science." In *The SAGE Handbook of Research Methods in Political Science and International Relations*, edited by Luigi Curini and Robert J. Franzese, 1053–78. 55 City Road: SAGE Publications Ltd. https://sk.sagepub.com/reference/the-sage-handbook-of-research-methods-in-political-science-and-ir/i8596.xml.

Ernst, Michael D. 2004. "Permutation Methods: A Basis for Exact Inference." *Statistical Science* 19 (4): 676–85. https://www.jstor.org/stable/4144438.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls

of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://academic.oup.com/pan/article/21/3/267/1579321/Text-as-Data-The-Promise-and-Pitfalls-of-Automatic.

Haxby, James V., M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. 2001. "Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex." *Science* 293 (5539): 2425–30. https://science.sciencemag.org/content/293/5539/2425.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael D. McDonald. 2007. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD 1990-2003*. Oxford: Oxford University Press.

Lappin, Shalom. 2017. "Formal Semantics." In *The Handbook of Linguistics*, edited by Mark Aronoff and Janie Rees-Miller. Hoboken: John Wiley & Sons. http://ebookcentral.proquest.com/lib/berkeley-ebooks/detail.action?docID=4822517.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20 (1): 78–91. https://www.cambridge.org/core/journals/political-analysis/article/coder-reliability-and-misclassification-in-the-human-coding-of-party-manifestos/145AC6C390225AB29DA0BBA99038E796.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." http://arxiv.org/abs/1301.3781.

Reimers, Nils, Philip Beyer, and Iryna Gurevych. 2016. "Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 87–96. Osaka, Japan: The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1009.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." http://arxiv.org/abs/1908.10084.

Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. 2020. "Adjusting for Confounding with Text Matching." *American Journal of Political Science* 64 (4): 887–903. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12526.

Rodriguez, Pedro L., Arthur Spirling, and Brandon M. Stewart. 2021. "Embedding Regression: Models for Context-Specific Description and Inference." *Working Paper*, June.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2021. "A Primer in BERTology: What We Know About How BERT Works." *Transactions of the Association for Computational Linguistics* 8 (January): 842–66. https://doi.org/10.1162/tacl_a_00349.

Ruedin, Didier, and Laura Morales. 2019. "Estimating Party Positions on Immigration: Assessing the Reliability and Validity of Different Methods." *Party Politics* 25 (3): 303–14. https://doi.org/10.1177/1354068817713122.

Werner, Annika, Onawa Lacewell, Andrea Volkens, Theres Matthiess, Lisa Zehnter, and Leila van Rinsum. 2021. *Manifesto Coding Instructions (5th Re-Revised Edition)*. Manifesto Project's Handbook Series. https://manifesto-project.wzb.eu/.

Wilkerson, John, and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20 (1): 529–44. https://doi.org/10.1146/annurev-polisci-052615-025542.