# Mining Meaning: Sentence Embedding and Semantic Similarity in the Analysis of Political Text

Marika Landau-Wells*

Working Paper, March 12 , 2022

**Abstract**

Text analysis is an increasingly valuable tool in the social sciences. Advances in Natural Language Processing (NLP) now allow researchers to make use of the semantic meaning of text, alongside standard frequency-based information. This research note demonstrates the value of using sentence-level (in contrast to word-level) NLP encoding models to characterize the semantic similarity of a set of texts. In three use-cases, I show how scholars can (1) validate qualitatively defined coding schemes using sentence-level encoding models and semantic similarity; (2) test for confounders in a corpus using the same method; and (3) extend a qualitative coding scheme to new data. Sentence-level embeddings and semantic similarity thus further expand the social science NLP toolkit and can complement other methods of text analysis.

Word count: 4,146

(including references)

*Travers Department of Political Science, University of California, Berkeley I would like to thank Rich Nielsen for helpful comments.

# Introduction

Since Grimmer and Stewart (2013) initially laid out the benefits (and pitfalls) of using text as data, scholars have embraced and extended methods developed in computer science and adjacent fields to better understand political texts and related phenomena. As a result, the use of computational tools to analyze political text has grown significantly (for reviews, see Wilkerson and Casas 2017; Benoit 2020; Chatsiou and Mikhaylov 2020).

The most common text analysis techniques rely on measuring the frequency of words or tokens, either as a primary quantity of interest or as a precursor to identifying the co-occurrences that define latent topics (e.g., Barberá et al. 2019; Lucas et al. 2015; Mueller and Rauh 2018; Quinn et al. 2010). With frequency-based measures, claims about the distinctions between texts are based on the (relative) presence or absence of specific terms within that corpus. The advantage of frequency-based measures is their simplicity; whether implemented over single words (or word-stems) or n-gram-length tokens, the basic principle is counting. What is lost with frequency-based measures, however, is semantic meaning at the word level (Benoit 2020).

Natural Language Processing (NLP) tools were developed, in part, to preserve semantic meaning using deep-learning techniques (Chatsiou and Mikhaylov 2020). A number of NLP deep-learning models use a process known as vectorization, encoding, or embedding. The embedding process transforms a piece of text (or an image or other object of interest) into a numeric vector. Initially, embedding models operated on words (e.g., Word2Vec Mikolov et al. 2013), transforming them into high-dimensional vectors in such a way that their proximity in the embedding space reflected some shared semantic properties. Terechshenko et al. (2020) demonstrate that embedding models generally outperform traditional models (e.g., random forests) in a classification task. Rodriguez and Spirling (2022) show that word embedding models can recover semantic relationships in political texts with performance similar to that of human raters.

Scholars have used word-level embeddings to estimate the political stances of Twitter users (Mebane Jr et al. 2018) and to compare their tweets on Brexit (Little et al. 2020); to infer ideological placement of parliamentary members (Rheault and Cochrane 2020); to assess the negativity in parliamentary speeches (Rudkowsky et al. 2018); to quantify change over time in gender stereotypes (Garg et al. 2018); and to understand the influence of partisanship on word usage (Rodriguez, Spirling, and Stewart 2021). Word-level embeddings can also be effectively averaged across longer segments, such as paragraphs, to generate embeddings for larger chunks of text (e.g., Le and Mikolov 2014).[1]

Word-level embeddings have their limitations, however. The basic unit of a thought expressed in text is more likely to be sentence-length, or longer. Therefore, treating the word as the primary unit of meaning limits what a model can capture. In addition, the task on which word-embedding models have been generally trained is not necessarily the task that researchers might wish to perform (e.g., next-word prediction for Word2Vec and Doc2Vec).

Reimers and Gurevych (2019) demonstrated that it is possible to encode much longer strings of text in a computationally efficient manner that allows sentences to be used as unit of embedding. Like other deep learning models, sentence-encoders can be trained and fine-tuned on different types of tasks to define what is considered proximate in the embedding space (e.g., matching questions and answers or titles and Reddit posts). There are now a variety of pre-trained embedding models that vary in their base architecture, training data, and fine-tuning (for examples, see https://huggingface.co/sentence-transformers).

In this note, I demonstrate how social scientists can use sentence-level embeddings derived from models pre-trained and fine-tuned on the task of semantic textual similarity (STS) to analyze properties of text when *meaning* rather than term-frequency is of primary interest. Meaning often matters more than term-frequency when researchers must rely on human

[1]Le and Mikolov (2014)'s approach generates a paragraph-level token that is learned in an unsupervised manner from the words within the paragraph, so it is not strictly an average. The paragraph-level token is still a function of word-level embeddings, however.

judgment of "what goes with what," rather than strict dictionaries, for evaluating texts. Here, I use sentence-level embeddings and a standard similarity metric in NLP – cosine similarity – across three use-cases: (1) validating a qualitatively-defined coding scheme; (2) testing for confounders within a corpus; and (3) extending an existing coding scheme to new texts. Each use-case reflects a task that scholars dealing with large amounts of text data might need to undertake. A worked example in R is provided in the Supplementary Materials.

## Sentence Embedding Models and Semantic Similarity

The degree of similarity between texts (e.g., official statements, manifestos, speeches, free response items) is often of fundamental interest to social scientists. Manuals for hand-coding text data represent attempts to define "what goes with what" along the dimension(s) a scholar cares about. Close reading of archival and other texts to demonstrate patterns and trends also draws on the human ability to detect subtle similarities and differences between texts. These tasks require qualitative assessments of texts that may share meaning without sharing words.

Sentence-level embedding models attempt to train machines to pick up on the same subtleties of meaning that humans do. The fundamental principle of these models is that sentences, like words, can be encoded in a high-dimensional space, such that proximity in that space between any two sentences reflects their similarity in some sense. In their initial formulation, Reimers and Gurevych (2019), note that there are number of use-cases for these embedding models, including semantic search, translation, and semantic textual similarity (STS). Indeed, sentence-encoding models can be pre-trained and fine-tuned for particular tasks or on multiple tasks.

I illustrate the advantage of sentence encoders in Table 1. The table shows the results of a common benchmarking task in model evaluation – correlating a model's estimates of similarity between sentence-pairs with ratings from a set of human coders. The sentence-pairs were

developed by (O'shea, Bandar, and Crockett 2014), which argues that a Pearson correlation is the appropriate standard for judging performance. As illustrated in the table, the two sentence encoders outperform both a word-level embedding model (with a difference that is significant at $p < 0.002$) and the conventional models originally evaluated by (O'shea, Bandar, and Crockett 2014). The sentences developed by O'shea, Bandar, and Crockett (2014) were not used in training either of the sentence encoder models. Nevertheless, the correlation between their similarity scores and human ratings exceeds 0.9.

Table 1: Benchmarking Task for Similarity Estimates

| Model | Type | Task | $\rho$ vs. Human Ratings |
|---|---|---|---|
| stsb-mpnet-base-v2 (Hugging Face) | Sentence encoder | STS-trained | 0.912 |
| all-mpnet-base-v2 (Hugging Face) | Sentence encoder | Multitask-trained | 0.906 |
| Average GloVe embeddings | Word encoder | None | 0.802 |
| O'Shea et al. (2014)'s LSA model | Latent Semantic Analysis | None | 0.693 |

Despite their equally strong performance in the benchmark task, Figure 1 shows why an STS-trained model is preferred over the multitask model.[2] The STS-trained model correctly assigns a relatively high degree of similarity between two sentences that share no words, but some meaning ("That is a happy person," "She smiles a lot"), while also detecting

[2]Note that the STS and multitask models in Figure 1 are based on the same fundamental architecture and differ only in their training and tuning.

dissimilarity where shared meaning is low ("That is a happy person," "Today is a sunny day").

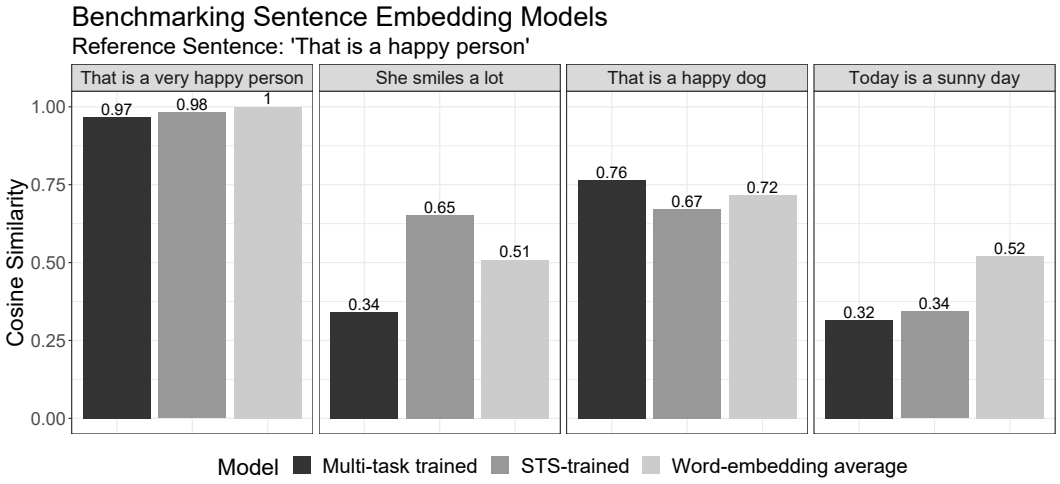


Figure 1: Sentence-Embedding Model Performance

The values in Figure 1 are cosine similarities, which can be calculated for any two vectors of embeddings $(\mathbf{A}, \mathbf{B})$ as their inner product:

$$\mathbf{S}(\mathbf{A}, \mathbf{B}) = \cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{A_i B_i}}{\sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{A_i})^{\mathbf{2}}} \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{B_i})^{\mathbf{2}}}} \tag{1}$$

Figure 2 illustrates how to scale up the type of comparison depicted in Figure 1 to an entire corpus. This can be done entirely in R using standard text analysis packages and an interface with a Python installation, such as the `reticulate` package (Ushey, Allaire, and Tang 2022).

In the remainder of this note, I demonstrate how researchers working with qualitatively-coded data (particularly small corpora that cannot support train/test splits for customizing models) can make use of pre-trained STS models. Specifically, I present three use-cases with `stsb-mpnet-base-v2` (Reimers and Gurevych 2019), the STS-trained model in Table 1 and Figure 1. The model embeds sentences (with a 75-word cut-point) into a 768-dimensional dense vector space. This model was initially trained on the Stanford Natural Language

Inference corpus (Bowman et al. 2015) and then fine-tuned using the STS benchmark dataset (Cer et al. 2017).

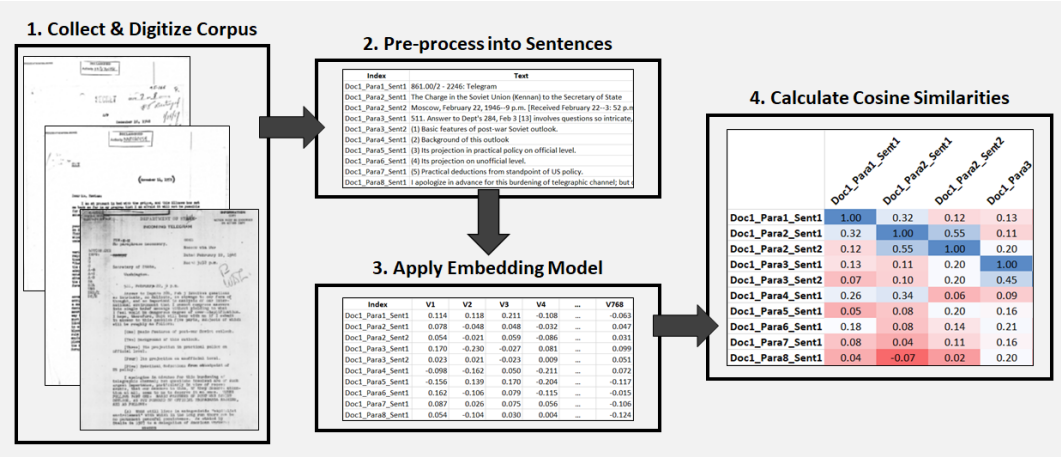


Figure 2: Raw Text to Similarity Workflow

# Use-Case 1: Validating Qualitative Coding Schemes

A common task for social scientists is the coding of textual material into categories. Sometimes the features that define categories are objectively obvious, indicated by the presence of certain terms, for example. Often, however, scholars want to organize a corpus into categories that defy dictionary-based methods of sorting. In these cases, deciding which texts are most alike along the dimension(s) of interest is left to the discretion of the researchers. It is common practice to employ more than one human coder in such cases and to report their level of agreement (i.e., intercoder reliability) across the corpus.

Here I illustrate an alternative to the multiple coder model (though these methods are by no means mutually exclusive) with a dataset of documents from the early Cold War concerning Communism. Each paragraph in the text 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 (coded 0).

A coding scheme such as this one implies that two texts which are assigned the same coded value should be more similar in semantic meaning to one another than to a text assigned to a different category. This *relative similarity* should hold even if the texts share no words in common.

If Category $\mathbf{X}$ is defined as a set of sentence-length texts whose embeddings are represented by vectors $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}$, then the *within-category similarity* of Category $\mathbf{X}$, i.e., $Sim_X$, can be represented in a square, symmetric similarity matrix:

$$Sim_X = \begin{bmatrix} S_{x1,x1} & S_{x1,x2} & \cdots & S_{x1,xn} \\ S_{x2,x1} & S_{x2,x2} & \cdots & S_{x2,xn} \\ \vdots & \vdots & \ddots & \vdots \\ S_{xn,x1} & S_{xn,x2} & \cdots & S_{xn,xn} \end{bmatrix} \tag{2}$$

*Within-category similarity*, $W_X$, is given by the mean of the row-wise averages of $Sim_X$, excluding the identity values. Thus, where $i$ and $j$ index rows and columns, respectively, so that $s_{ij}$ denotes the elements of $Sim_X$:

$$W_X = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} s_{ij} \tag{3}$$

If Category $\mathbf{Y}$ is defined as a set of sentence-length texts whose embeddings are represented by vectors $\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n}$, then the *between-category similarity* for Categories $\mathbf{X}$ and $\mathbf{Y}$, $Sim_{XY}$, can be represented in a symmetric similarity matrix as well:

$$Sim_{XY} = \begin{bmatrix} S_{x1,y1} & S_{x1,y2} & \cdots & S_{x1,ym} \\ S_{x2,y1} & S_{x2,y2} & \cdots & S_{x2,ym} \\ \vdots & \vdots & \ddots & \vdots \\ S_{xn,y1} & S_{xn,y2} & \cdots & S_{xn,ym} \end{bmatrix} \tag{4}$$

Using the same notation as Equation 3, *between-category similarity*, $B_{XY}$, is given by the mean of the row averages of $Sim_{XY}$, which is an $n$ row by $m$ column matrix.

$$B_{XY} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m} s_{ij} \tag{5}$$

The *relative similarity* of Category **X** with respect to Category **Y** is thus a difference score:

$$WB_{XY} = W_X - B_{XY} \tag{6}$$

If the value $WB_{XY}$ is positive, then Category **X** texts are discriminable from Category **Y** texts (i.e., they are more 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 cosine similarity metric itself is somewhat arbitrarily scaled, because it is a function of the chosen embedding model, the $WB_{XY}$ value is directly interpretable. Significance testing can be performed using the $WB_{XY}$ statistic, since the null hypothesis of no semantic meaning differences between categories is represented by $WB_{XY} = 0$. As the assumption of normality is quite strong with respect to cosine similarity, non-parametric methods are preferred for hypothesis testing.

Panel A of Figure 3 shows the average semantic similarity within each category (the diagonal) and across categories (the off-diagonals). All within-category values are distinguishable at $p < 0.002$.[3] As noted above, significance testing can be performed on the $WB_{XY}$ statistic across all category pairings, the results of which are shown in Panel B of Figure 3. As Panel B shows, all observed *relative similarities* lie outside the null distribution, which suggests that the qualitatively-defined hand-coding scheme captures true differences in semantic meaning that can be recognized by an independent coder, i.e., the STS embedding model. In the Supplementary Materials, I illustrate the relatively poor performance of embedding models not trained on STS for this use-case, which shows the importance of pre-training on the task of interest.

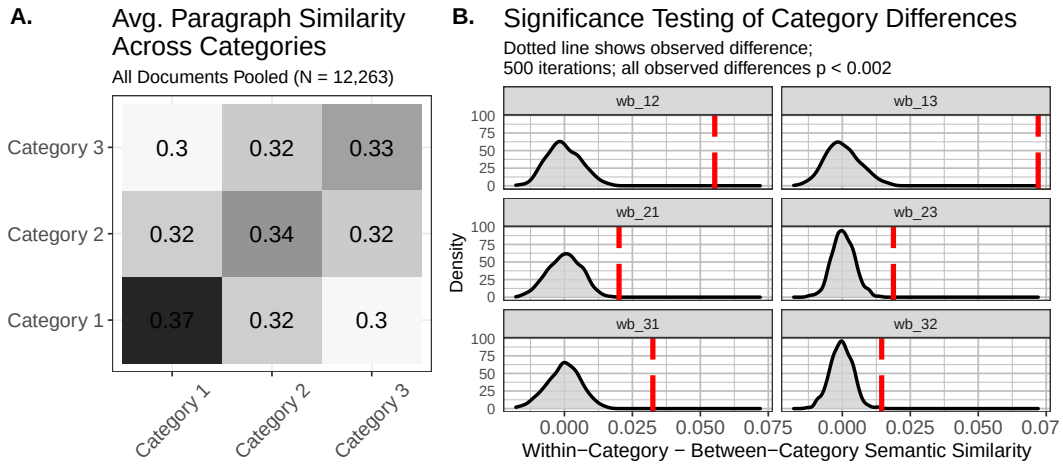


Figure 3: Coding Category Similarity

# Use-Case 2: Testing for Confounders

As Use-Case 1 demonstrated, semantic similarity is a relative property of any two texts once they have been transformed into a shared embedding space. 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

[3] An alternative method that first calculates each statistic at the author-level is discussed in the Appendix. The findings hold either way.

confounder. Some confounders might affect the frequency with which texts are produced, e.g., Roberts, Stewart, and Nielsen (2020)'s estimate the unseen effects of prior censorship on the number of subsequent social media posts. But confounders might also affect semantic content. In the case of the Cold War Corpus, we 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. Using the same metric as Use-Case 1, $WB_{XY}$, and permutation to derive a null distribution, we can test whether or not secrecy status matters for the semantic content of texts within the corpus. Figure 4 shows the results of 500 permutations of the three secrecy status labels (Top Secret, Off-the-Record, Never Classified) on the threat content of each document, calculating the $WB_{XY}$ for each category-pair. As we would expect, these null distributions are centered on zero. The vertical lines indicate the observed value in the corpus. For each category-pair, the observed value lies well within the null distribution. From this, we can conclude that the threat content of Top Secret documents is not semantically distinguishable from documents designated for public consumption (i.e., Never Classified or Off-the-Record). This gives some confidence that the threat-related content of these documents is not a function of their intended audience. In principle, any categorical potential confounder can be tested this way.

## Use-Case 3: Extending a Coding Scheme to New Text

Scholars often confront a corpus that extends beyond the limits of reading, which poses a challenge for qualitative coding schemes. Semantic similarity offers a way in which the coding scheme applied to one corpus can be extended to another, or from a random subset of texts to a whole corpus. In effect, the high-dimensional embedding space acts as a bridge that allows us to link texts in a consistent way. This is one way around the transfer learning problem that arises using task-specific models (Cer et al. 2018).

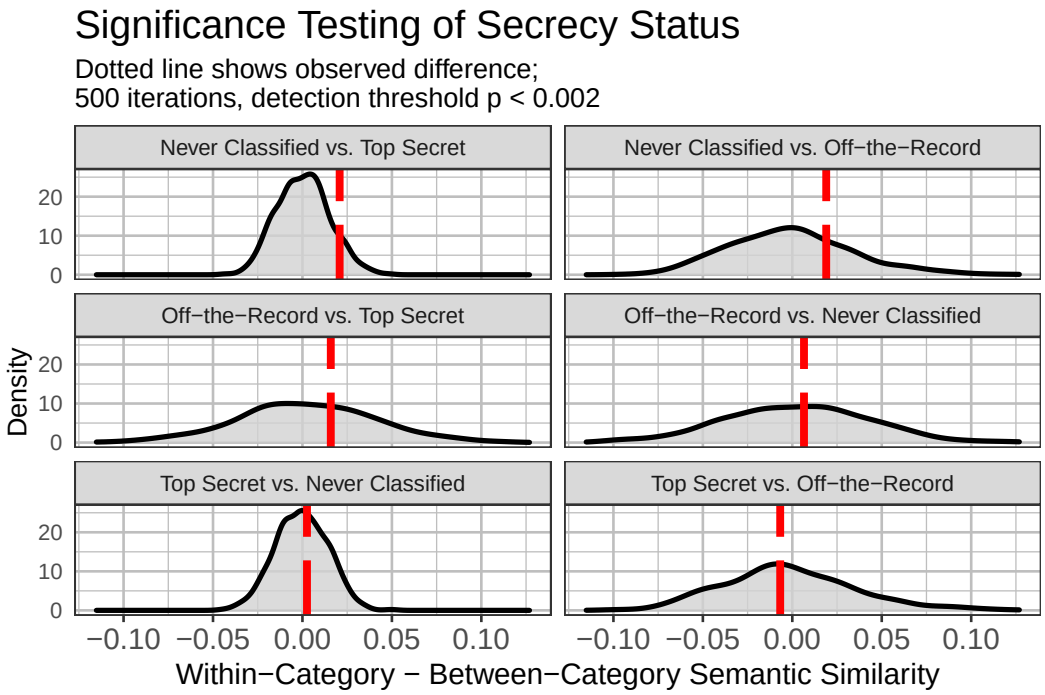


Figure 4: Testing for Confounders

To extend the coding scheme developed in the Cold War corpus to a new body of text – speeches given by President George W. Bush – I use a rank-based classification algorithm where semantic similarity defines the ranking system. First, the new corpus is fed into the same embedding model as the original corpus. Then, instead of calculating cosine similarity within the new corpus, I calculate the similarity between every document of the new corpus and the original. I then identify the most similar documents in the original corpus (Top 5 in this case) for every document in the new corpus. The classification algorithm then codes the new text as Not Threat Relevant if all matches in the original corpus are 0 and takes the most similar coding (Category 1, 2, or 3) otherwise.

To examine the face validity of this approach, we can look for patterns in the new corpus that we are already aware of. For instance, Figure 5 shows that George W. Bush's speeches contained more threat-related content after September 11, 2001 than before – something which is hardly surprising, but gives us some confidence in the extension of the coding scheme.

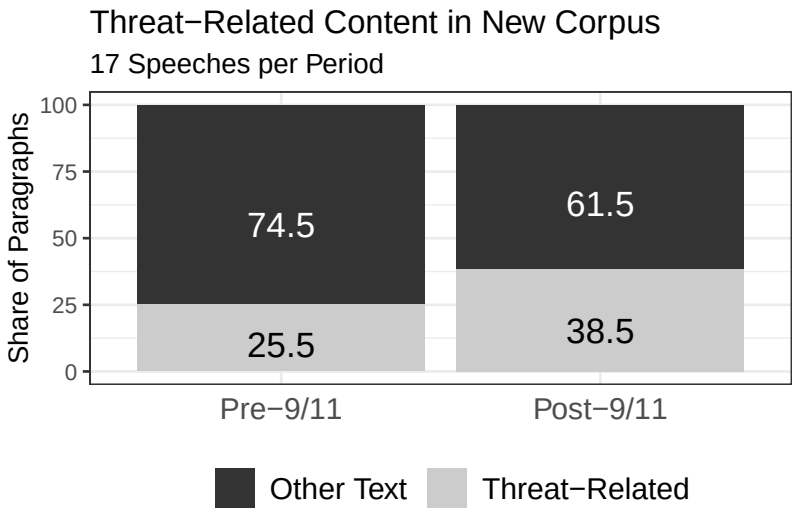


Figure 5: Coding A New Corpus

To enhance confidence in the assigned labels, the researcher can (1) examine the labels (or some random sample) manually; (2) use unsupervised tools such as topic-modelling (e.g., Roberts et al. 2014) or relative keyness (e.g., Wiedemann 2013) to examine what terms appear uniquely in the categories within the new corpus; or (3) use a "Turing assessment" (Rodriguez and Spirling 2022) to evaluate the alignment of the ranking algorithm with human judgment. As Tables 2 and 3 show, the top terms characterizing the categories in the new corpus do not overlap with the original corpus from which the coding was sourced. This is consistent with a classification model that derives coding from a shared space of meaning (defined by the embedding model), rather than a shared dictionary of terms.

Table 2: New Corpus: Most Distinct Terms by Category
($p < 0.001$)

| | Existential Harm | Loss | Contamination |
|---|---|---|---|
| 1. | disarm | achieve | hate |
| 2. | threat | freedom | evil |
| 3. | weapons_of_mass_destruction | oceans | patriot |
| 4. | weapons | regimes | country |
| 5. | allies | peace | frivolous |

Table 3: Original Corpus: Most Distinct Terms by Category ($p < 0.001$)

| | Existential Harm | Loss | Contamination |
|---|---|---|---|
| 1. | attack | domination | propaganda |
| 2. | striking | soviet | infiltration |
| 3. | soviet-manufactured | imperialism | unions |
| 4. | recalling | conquest | penetration |
| 5. | catastrophe | control | fertile |

# Conclusion

NLP models are continually under development in computer science and elsewhere, creating opportunities for social scientists to mine more meaning from text data. In this note, I have shown that sentence-level embeddings and semantic similarity analysis can be used to: (1) validate qualitatively-defined coding schemes; (2) test for confounders that might systematically affect the distribution of semantic meaning across a corpus; and (3) extend qualitative coding schemes to large amounts of new texts. In addition to the STS-trained model tested across these three use-cases, other model variants exist trained and fine-tuned on other tasks that might also be of interest to researchers (e.g., matching questions to answers). In sum, these tools provide new ways of assessing meaning-based relationships between texts – relationships that are often at the heart of our coding and classification schemes and of our theories.

# Appendix

An R script that generates sentence embeddings and similarity measures with a worked example is available on github.

## Significance Testing for Coding Category Discrimination

There are two ways of measuring category discrimination in the Cold War corpus. The first considers model performance over all texts, regardless of authorship (the Pooled approach). This is illustrated in the main text. The second considers author-level averages (the Individual-Level approach), which reduces the influence of individuals with relatively more text in the corpus. In this case, either approach returns the same result.

With the Individual-level averages (N = 22), I use a non-parametric test (paired Wilcoxon Signed Rank test) to test the one sided hypothesis that Within-Category similarity, $W_X$, is always greater than Between-Category similarity, $B_{XY}$. For Category 1, $WB_{12}$ V = 40, p-value = 0.02; $WB_{13}$ V = 45, p-value = 0.002. For Category 2, $WB_{21}$ V = 110, p-value = 0.01; $WB_{23}$ V = 157, p-value = 0.03. For Category 3, $WB_{31}$ V = 74, p-value = 0.02; $WB_{32}$ V = 146, p-value = 0.003.

## Model Performance

The task reported in Use-Case 1 was run on a number of different sentence-transformers. I report the results here to illustrate the relatively superior performance of models trained and fine-tuned on Semantic Textual Similarity (STS). These models differ in their initial pre-training foundations (i.e., MPNet (Song et al. 2020) versus RoBERTa (Liu et al. 2019)), in their initial training data (i.e., Stanford's NLI dataset (Bowman et al. 2015) versus the unsupervised web data used by Google's Universal Sentence Encoder (USE) (Cer et al. 2018)) and in whether or not they were fine-tuned. The results in Table 4 are for performance using the Individual-Level approach described in the previous section as this is the harder task.

Performance is defined as the number of categories that can be distinguished (maximum possible score = 3).

Table 4: Category Discrimination Task Performance

| Model | Training Data | Fine-Tuning | Dimensions | Performance |
|---|---|---|---|---|
| stsb-mpnet-base-v2 | Stanford NLI | STS benchmark dataset | 768 | 3/3 |
| stsb-roberta-base-v2 | Stanford NLI | STS benchmark dataset | 768 | 3/3 |
| all-mpnet-base-v2 | Multiple sentence pairing tasks | None | 768 | 2/3 |
| all-MiniLM-L12-v2 | Multiple sentence pairing tasks | None | 384 | 1/3 |
| all-MiniLM-L6-v2 | Multiple sentence pairing tasks | None | 384 | 1/3 |
| all-distilroberta-v1 | Multiple sentence pairing tasks | None | 768 | 0/3 |
| USE-DAN | Unsupervised web data + SNLI | None | 512 | 2/3 |
| USE-Transformer | Unsupervised web data + SNLI | None | 512 | 0/3 |
| average-word-embeddings_glove.6B.300d | Wikipedia + Gigaword corpora | None | 300 | 0/3 |
| average-word-embeddings_levy_dependency | Wikipedia | None | 300 | 0/3 |

# References


Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. 2019. "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data." *American Political Science Review* 113 (4): 883–901. https://doi.org/10.1017/S0003055419000352.

Benoit, Ken. 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. 55 City Road: SAGE Publications Ltd. https://sk.sagepub.com/reference/the-sage-handbook-of-research-methods-in-political-science-and-ir/i4365.xml.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. "A Large Annotated Corpus for Learning Natural Language Inference." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. "SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation." *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 1–14. http://arxiv.org/abs/1708.00055.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, et al. 2018. "Universal Sentence Encoder for English." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–74. Brussels, Belgium: Association for Computational Linguistics. https://aclanthology.org/D18-2029.

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.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115 (16): E3635–44. https://www.pnas.org/content/115/16/E3635.

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.

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *Proceedings of the 31st International Conference on Machine Learning*,

1188–96. PMLR. https://proceedings.mlr.press/v32/le14.html.

Little, Claire, David Mclean, Keeley Crockett, and Bruce Edmonds. 2020. "A Semantic and Syntactic Similarity Measure for Political Tweets." *IEEE Access* 8: 154095–113. https://doi.org/10.1109/ACCESS.2020.3017797.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." July 26, 2019. http://arxiv.org/abs/1907.11692.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–77. https://www.cambridge.org/core/journals/political-analysis/article/computerassisted-text-analysis-for-comparative-politics/CC8B2CF63A8CC36FE00A13F9839F92BB.

Mebane Jr, Walter R, Patrick Wu, Logan Woods, Joseph Klaver, Alejandro Pineda, and Blake Miller. 2018. "Observing Election Incidents in the United States via Twitter: Does Who Observes Matter?" In *Annual Meeting of the Midwest Political Science Association, Chicago*.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." September 6, 2013. http://arxiv.org/abs/1301.3781.

Mueller, Hannes, and Christopher Rauh. 2018. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review* 112 (2): 358–75. https://www.cambridge.org/core/journals/american-political-science-review/article/reading-between-the-lines-prediction-of-political-violence-using-newspaper-text/4EABB473AFE18F157EEDE4339F34ABB0.

O'shea, James, Zuhair Bandar, and Keeley Crockett. 2014. "A New Benchmark Dataset with Production Methodology for Short Text Semantic Similarity Algorithms." *ACM Transactions on Speech and Language Processing* 10 (4): 19:1–63. https://doi.org/10.1145/2537046.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–28. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2009.00427.x.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." August 27, 2019. http://arxiv.org/abs/1908.10084.

Rheault, Ludovic, and Christopher Cochrane. 2020. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* 28 (1): 112–33. https://www.cambridge.org/core/journals/political-analysis/article/word-embeddings-

for-the-analysis-of-ideological-placement-in-parliamentary-corpora/017F0CEA9B3DB6E1B94AC36A509A8A7B.

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.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12103.

Rodriguez, Pedro L., and Arthur Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 84 (1): 101–15. https://www.journals.uchicago.edu/doi/full/10.1086/715162.

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

Rudkowsky, Elena, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. 2018. "More Than Bags of Words: Sentiment Analysis with Word Embeddings." *Communication Methods and Measures* 12 (2-3): 140–57. https://doi.org/10.1080/19312458.2018.1455817.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. "MPNet: Masked and Permuted Pre-training for Language Understanding." In *Advances in Neural Information Processing Systems*, 33:16857–67. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.

Terechshenko, Zhanna, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2020. "A Comparison of Methods in Political Science Text Classification: Transfer Learning Language Models for Politics." SSRN Scholarly Paper ID 3724644. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3724644.

Ushey, Kevin, JJ Allaire, and Yuan Tang. 2022. *Reticulate: Interface to 'Python'*. Manual.

Wiedemann, Gregor. 2013. "Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences." *Historical Social Research / Historische Sozialforschung* 38: 332–57. https://www.jstor.org/stable/24142701.

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.