Journal of Computational Social Science (2021) 4:231–242
https://doi.org/10.1007/s42001-020-00075-8

RESEARCH ARTICLE

Integrating semantic directions with concept mover's distance to measure binary concept engagement

Marshall A. Taylor[1] · **Dustin S. Stoltz**[2]

Received: 8 March 2020 / Accepted: 26 June 2020 / Published online: 14 July 2020

Abstract

In an earlier article published in this journal ("Concept Mover's Distance", 2019), we proposed a method for measuring concept engagement in texts that uses word embeddings to find the minimum cost necessary for words in an observed document to "travel" to words in a "pseudo-document" consisting only of words denoting a concept of interest. One potential limitation we noted is that, because words associated with opposing concepts will be located close to one another in the embedding space, documents will likely have similar closeness to starkly opposing concepts (e.g., "life" and "death"). Using aggregate vector differences between antonym pairs to extract a direction in the semantic space pointing toward a pole of the binary opposition (following "The Geometry of Culture," *American Sociological Review*, 2019), we illustrate how CMD can be used to measure a document's engagement with binary concepts.

Keywords Concept mover's distance · Geometry of culture · Word embeddings · Text analysis · Cultural sociology · Natural language processing

Introduction

In an article published in the *Journal of Computational Social Science* [20], we presented a method for measuring concept engagement in texts. The method—called concept mover's distance (CMD)—is an extension of word mover's distance (WMD; [11]) that uses word embeddings and the earth mover's distance algorithm [2, 17] to find the minimum cost necessary for words in an observed document to "travel" to words in a pseudo-document—a document consisting only of words denoting a concept of interest.

✉ Marshall A. Taylor
mtaylor2@nmsu.edu

[1] Department of Sociology, New Mexico State University, Las Cruces, USA

[2] Department of Sociology and Anthropology, Lehigh University, Bethlehem, USA

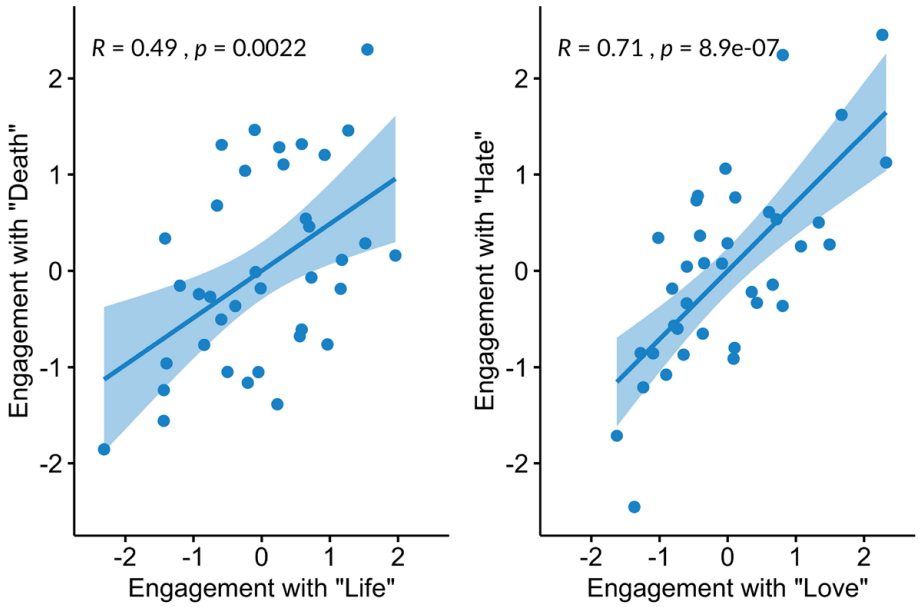

Fig. 1 Correlations between CMD for 'Life' and 'Death' (left panel) and 'Love' and 'Hate' (right panel) in Shakespeare's First Folio. Pearson's correlation coefficients reported. Bands are 95% confidence intervals. For data preprocessing steps taken, see [20, pp. 299–300]. All plots in this paper were made with some combination of ggplot2 [22] and ggpubr [9].

We demonstrated the utility of CMD using a range of publicly available corpora, showing how CMD can be used to chart the emergence of concepts over time, map concept engagement onto external events, and measure engagement with more specified or compound concepts. We also illustrated how CMD is robust to sparse term removal and can even be used when terms denoting a concept of interest are absent from the corpus.

One limitation we noted, however, relates to binary concepts—concepts that oppose one another in some meaningful way, such as "life" and "death," "good" and "evil," or "sacred" and "profane." The issue arises not directly from CMD, but as a downstream consequence of the fact that CMD measures distances using word embeddings.

Since, in most standard embedding models, similarity metrics between word embedding vectors quantify the extent to which two words are used in contextually similar ways (but not necessarily in the form of actual co-occurrences), words that are used in opposition to one another necessarily occupy similar positions in the n-dimensional embedding space precisely because they are mutually oriented toward a shared (though diametrically opposed) meaning [8]. This is a widely known issue in distributional semantics (e.g., [18]) and word embedding research (e.g., [14]).

The consequence of this for CMD is that documents will likely have similar closeness to starkly opposing concepts. Using CMD with Shakespeare's First Folio, for example, we find reasonably strong and positive correlations between a play's conceptual engagement with "life" and "death" (0.49) and with "love" and "hate" (0.71; see Fig. 1) [20].

The question, then, is this: How can CMD be used to measure the extent to which a document is close to a particular pole of a binary opposition? In this note, we propose one such method from recent work on word embeddings in cultural sociology. Specifically, we incorporate work from Kozlowski et al. [10] on cultural dimensions into the CMD workflow to show how CMD can be used to measure a document's engagement with binary concepts.

In what follows, we outline Kozlowski, Taddy, and Evans' work and show how their procedure for defining the "semantic direction" of a cultural dimension within an embedding space (and similar procedures) can be easily integrated with CMD to establish the "pole" of binary concepts. We then present an illustration, using CMD to measure engagement with the "liberal vs. conservative" binary in all U.S. State of the Union addresses from 1790 to 2019. Next, we compare how well the "death vs. life" binary predicts actual deaths in Shakespeare's plays as compared to the the single terms "death" and "life."

Cultural dimensions and the "geometry of culture"

Kozlowski, Taddy, and Evans use word embeddings trained at different time periods in an English-language corpus to estimate changes relative to key cultural dimensions, which they define as relational meaning structures that "individuals use in everyday life to classify agents and objects in the world" [10, p. 911]. Specifically, they use the example of social class to show how shared meanings of affluence, cultivation, education, and status remained fairly stable across the twentieth century, although cultivation has grown less associated with affluence while education has grown more associated.

Building off the well-known vector-offset method [15], the authors are able to locate the positions of "cultivation" and "education" relative to "affluence," for example, by defining a "semantic direction" (i.e. a one-dimensional subspace) within the word embedding space using antonym pairs. Mathematically, this semantic direction, $\mathbf{d}$, is the mean of a set of word vector differences between a collection of antonym word pairs [10, p. 918]:

$$\mathbf{d} = \frac{\sum_{p}^{P} (\mathbf{p_1} - \mathbf{p_2})}{P}, \tag{1}$$

where p is an antonym pair in the total set of P-relevant antonym pairs, and $\mathbf{p_1}$ and $\mathbf{p_2}$ are the vector representations of the two words in antonym pair p, and $\mathbf{d}$ points toward $\mathbf{p_1}$ and away from $\mathbf{p_2}$ (see the Appendix for alternative procedures to estimate semantic directions).

To make this more intuitive, consider "affluence" as a key cultural dimension of the concept of social class [10, p. 912–3]. Say we are interested in which sports are associated with upper class and which are associated with working class. We could measure the distance (specifically, we will use cosine similarity which is bound between 1, for exactly the same, and −1, for exactly the opposite) between "tennis," "bowling," or "soccer," on the one hand, and "rich," on the other (see Fig. 2,

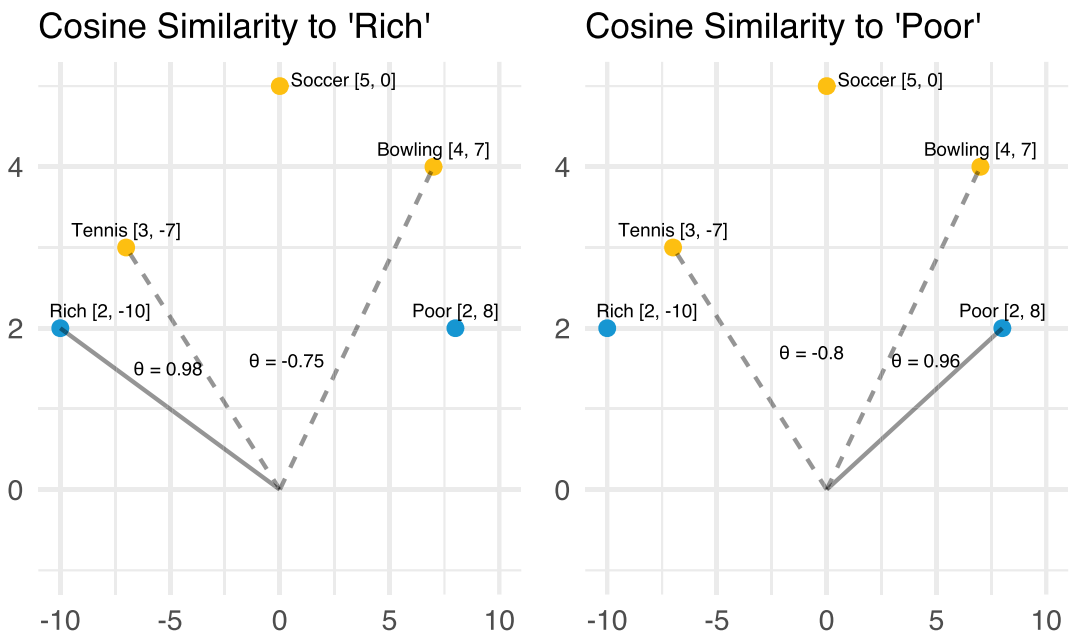

Fig. 2 Example of cosine similarities between word vectors. These are hypothetical two-dimensional word vectors

left panel). Let us say we find "tennis" is closer to "rich" than either "bowling" or "soccer."

We can interpret this in two ways: (1) bowling and soccer are less associated with *class* than is tennis, or (2) bowling and soccer are less associated with the *upper* class than is tennis. We can adjudicate between these two possible interpretations by bringing in the other pole of the affluence dimension: "poor." We might find that "bowling" is not only farther from "rich" than is "tennis," but also closer to "poor" (see Fig. 2, right panel).

Instead of combining these two cosine similarities after the fact, we can simplify this process by first subtracting the vector for "poor" from the vector for "rich," thus deriving a direction in the semantic space pointing toward affluence and away from non-affluence. We would then get a single cosine similarity between our term of interest and the vector for this semantic direction.

Importantly, this new vector is not a word vector per se, but rather a location in the multidimensional semantic space created by the word vectors, which some refer to as a relation vector [6]. For example, Fig. 3 includes the same sports terms, but now also includes a new point corresponding to the rich "pole" of the cultural dimension of affluence created by subtracting the vector for "poor" from the vector for "rich." The cosine similarity between "bowling" and this "pole" is about −0.87,[1] or highly opposed.

Furthermore, since the terms "rich" and "poor" are also associated with concepts other than affluence, we can further specify the semantic direction corresponding

[1] Which we could get by subtracting the cosine similarity between "bowling" and "rich" (−0.754) from the cosine similarity between "bowling" and "poor" (0.962), and dividing by two.

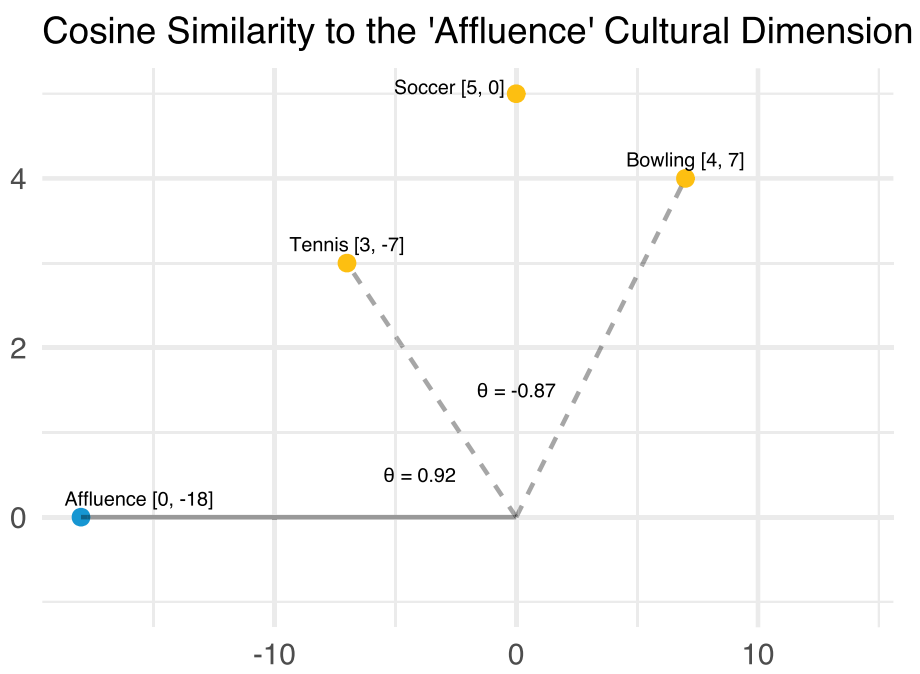

Fig. 3 Cosine similarities between a cultural dimension and each word vector in a document. These are hypothetical two-dimensional word vectors

to this cultural dimension using more antonym pairs. For example, to quantify their "affluence" dimension of social class, Kozlowski, Taddy, and Evans collected a total of 44 antonym pairs ranging from "rich–poor" to "opulence–indulgence" [10, p. 935], found the differences between the word vectors representing each word in the pair using subtraction, and then took the mean of those 44 vector differences to arrive at their "affluence" dimension (see Eq. 1). They used similar procedures to derive cultural dimensions of race, gender, morality, and status.

When an analyst is interested in measuring how texts engage with a concept that has a culturally obvious opposing concept (e.g., "love" vs "hate," "good" vs "bad"), using this procedure will allow the analyst to measure whether the texts are engaging more with one pole or the other.

Integrating semantic directions with concept mover's distance

Measuring a document's distance from a pole of a semantic direction with CMD is straightforward. In the original CMD paper, we proposed building pseudo-documents which could be filled with one or more words denoting a key concept. For example, let us say we want to measure the distance (technically, the similarity since CMD assigns higher values to documents with more concept engagement) between the sentence "The band gave a concert in Japan" and the concept of music. We would create a basic document-term matrix (after removing common function words, "the," "a," and "in") and then simply add a row for a pseudo-document only consisting of the word "music" (see Table 1).

Table 1 DTM with real document (row 1) and the pseudo-document for 'music' (row 2)

	Band	Gave	Concert	Japan	Music
Doc_{r1}	1	1	1	1	0
Doc_p	0	0	0	0	**1**

Note: Bold text indicates the token added to the pseudo-document

Table 2 DTM with real document (row 1) and the pseudo-document for a cultural dimension (row 2)

	Band	Gave	Concert	Japan	$D_{\text{affluence}}$
Doc_{r1}	1	1	1	1	0
Doc_p	0	0	0	0	**1**

Note: Bold text indicates the token added to the pseudo-document

The underlying algorithm [20, pp. 296–9] then calculates the minimum cost of moving all the words in each document to the positions of those one or few words in the pseudo-document. The cost is a function of the distances, as defined by the word embedding vectors associated with each word, as well as the count of each unique word. A document is said to engage more with a concept if the cost is low.

After using the method described in the previous section to derive the vector for a semantic direction of interest, the only alterations for CMD are to (1) add that vector to the word vector matrix, (2) add a new column to the document-term matrix (DTM) corresponding to the new semantic direction, and then (3) define a pseudo-document (i.e., a new row in the DTM) consisting only of a 1 in the cell where the row intersects with the new semantic direction column and 0s elsewhere. We then measure the cost it would take to move all the words in an observed document to that pseudo-document.

Let us say we want to measure the distance between the sentence "The band gave a concert in Japan" and the cultural dimension of affluence described previously. We add a column corresponding to the vector derived from subtracting "poor" from "rich"—here $D_{\text{affluence}}$. Next, we add a row for the pseudo-document which consists of all zeros except for a one in the same column as $D_{\text{affluence}}$ (see Table 2). (In our specific implementation, the name of this column corresponds to our new "pole" vector that is added to the word embedding matrix.) We would then add the minimum cosine similarities between the pole and each word in the document (weighted by the count of each term) to arrive at an overall concept engagement score (see Fig. 4).[2]

[2] See [20, pp. 296–9] and [11] for more detailed discussions of the underlying algorithm. Several teams have found computationally efficient methods of solving the transportation problem and our method now incorporates linear complexity relaxed word mover's distance [2], as implemented in the text2vec package [19].

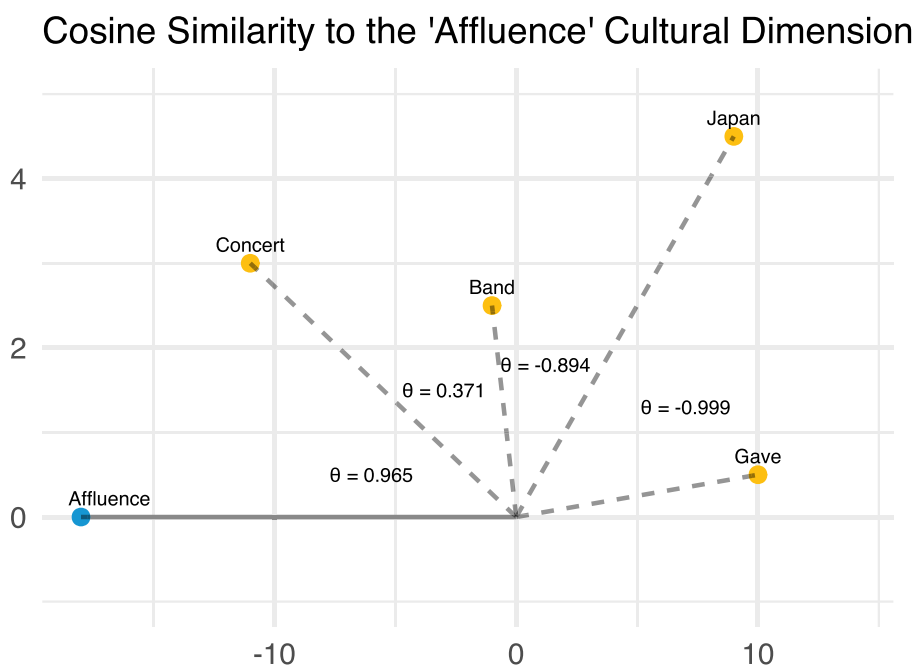

Fig. 4 Example of cosine similarities between word vectors and a cultural dimension. These are hypothetical two-dimensional word vectors. To get an aggregate engagement score for each document, the minimum cosine similarities between the pole and each word in the document are added and weighted by the count of each term. In the final measure, higher scores are interpreted as more engagement with a concept

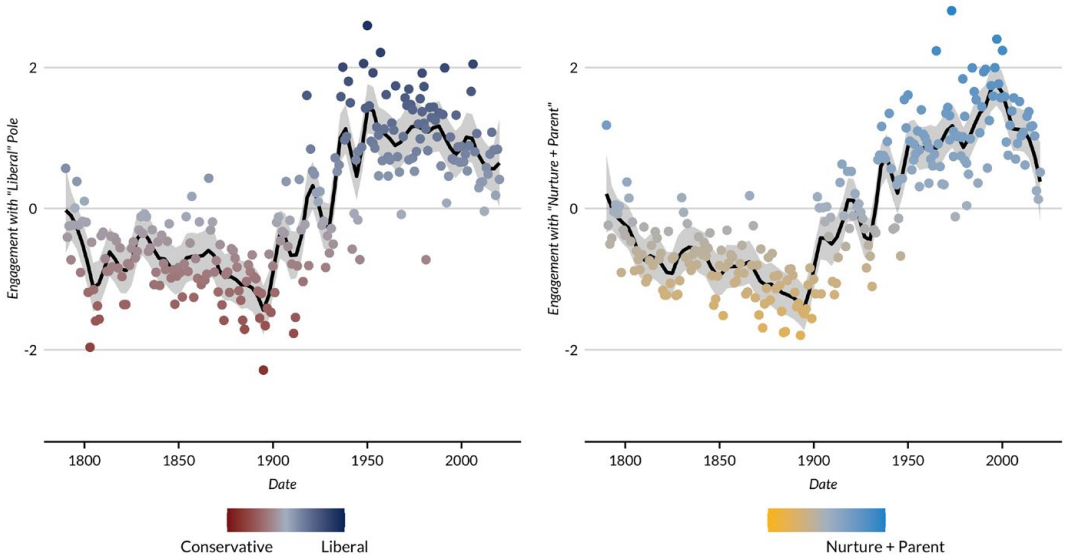

Fig. 5 Engagement with the "Liberal" pole of the "political ideology" dimension (left) and "nurturing parent" (right) in U.S. State of the Union Addresses, 1790–2019. Lines are smoothed with LOESS. Gray bands are 95% confidence intervals

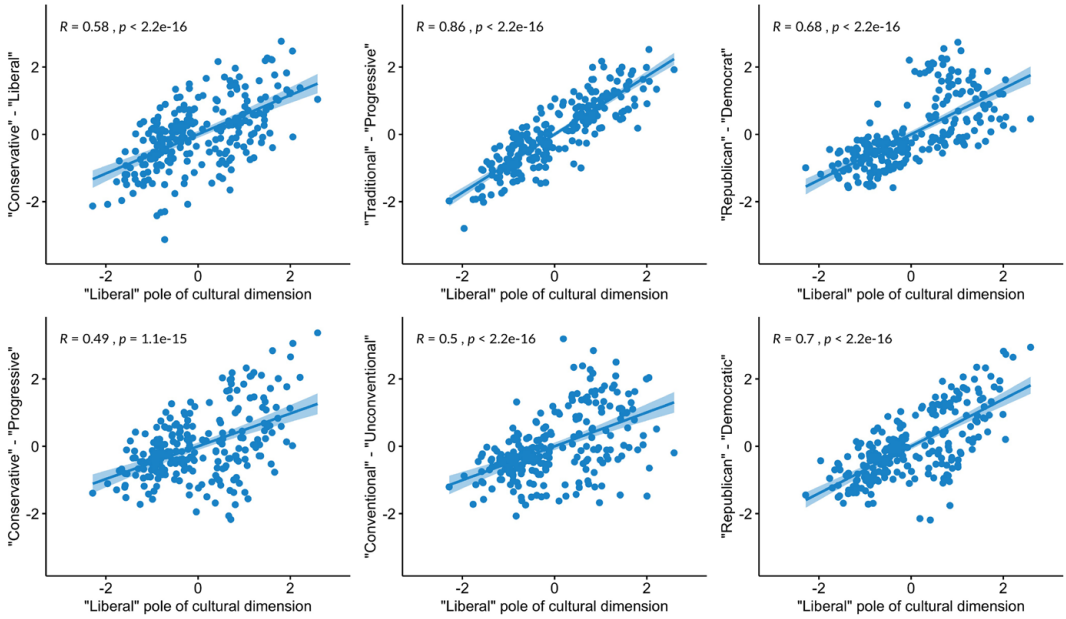

Fig. 6 Engagement with the "liberal" pole of the "political ideology" dimension by component antonym pairs. Points are individual State of the Union addresses

Demonstration: "liberal" vs. "conservative" in U.S. presidential discourse

Using the U.S. State of the Union Addresses [23], we measure each speech's engagement with the liberal–conservative binary (Fig. 5). To derive this cultural dimension of American politics, we use the following six antonym pairs: liberal–conservative, progressive–traditional, progressive–conservative, democrat–republican, democratic–republican, and unconventional–conventional. Following the same methods used in [20],[3] we use the fastText pre-trained English word embeddings.

In the previous paper, we showed how close each speech was to ideal-typical models of the family, building off Lakoff [12]. Specifically, he argues that the "nurturing parent" is the metaphor by which liberal-leaning people understand politics. In the original paper, we found that engagement with the "nurturing parent" had increased over the last century, gaining dominance over the conservative "strict father" model, and declining only in recent years. Using CMD with our cultural dimension of political ideology, we find engagement with the "nurturing parent" tracks rather well with each speech's engagement with the "liberal" pole (with a correlation of 0.89). Overall, this is commensurate with Lakoff's argument.

To see whether one antonym pair accounts for most of the orientation of the semantic direction, we compared engagement with each component antonym pair against engagement with the "liberal" pole. We find that each pairing is highly

[3] There are two differences in the SOTU example from [20]. First, there are 241 speeches in this analysis while there were 239 in [20]. Second, non-ASCII characters were removed as part of the processing procedure in the present analysis.

Table 3 Log count estimates of body counts by conceptual engagement

	Model 1	Model 2	Model 3
"Death" pole	0.926*** (0.171)		
"Death" word		0.652*** (0.172)	
"Life" word			−0.036 (0.201)
Constant	0.958*** (0.158)	1.114*** (0.171)	1.294*** (0.198)
N	37	37	37
AIC	162.36	173.75	184.83
−2LL	156.357	167.750	178.828
θ (overdispersion)	2.71 (1.38)	1.444 (0.591)	0.847 (0.285)
$\chi^2_{\text{nb-Poisson}}$ (1 *df*)	13.415***	27.815***	59.198***

Negative binomial log count estimates reported. Standard errors in parentheses. The θ is the overdispersion parameter, as calculated with the glm.nb function in the MASS R package [21]. $\chi^2_{\text{nb-Poisson}}$ is a chi-square likelihood ratio test assessing whether or not the overdispersion parameter adds a statistically significant improvement in model fit over the Poisson model

***$p < 0.001$, **$p < 0.01$ (two-tailed tests)

correlated, suggesting that no one single antonym pair is driving the orientation (see Fig. 6).

Demonstration: predicting deaths in Shakespeare's plays

In the original paper, we found a moderately strong correlation between engagement with the concept of "death" and the number of deaths in Shakespeare's plays from the First Folio [16], and we speculated that this may be downwardly biased due to the binary concept problem. Here, we compare how engagement with the terms "death," "life," and the cultural dimension of death–life predicts the body count in Shakespeare's plays. To derive the semantic direction for our cultural dimension, we used the following antonym pairs: death–life, casualty–survivor, demise–birth, dying–living, and fatality–endure.

Using negative binomial regression (see Table 3), we find that a play's closeness to the "death" pole of the cultural dimension had a positive and statistically significant relationship with the number of actual deaths in the plays, more so than the closeness to the word "death" on its own. Specifically, for a play that engages with the death pole of the death–life cultural dimension one standard deviation above the mean level of engagement, the predicted number of deaths is about 6.58 ($e^{0.958+0.926}$). This is in contrast to a play that engages the concept of "death" denoted by a single term one standard deviation above the mean, which translates to about 5.85 deaths ($e^{1.114+0.652}$). The death pole model has the best model fit, with AIC = 162.36 (versus AIC = 173.75 for just the term "death"). The "life" concept on its own had no statistically significant association with body count ($z = -0.180$; $p = 0.857$).

Conclusion

In the original CMD paper [20], we briefly discussed the binary concept problem: documents will be close to both polarities of concepts that oppose one another in some meaningful way, such as "life" and "death," "good" and "evil," or "sacred" and "profane," even if the analyst only wishes to measure engagement with one pole or the other of this binary concept. The issue arises because words that are used in opposition to one another will be placed in similar positions in the embedding space since they are both oriented toward a shared meaning, despite their opposing valences.

In this paper, we demonstrated one technique for overcoming this problem when using word embeddings to measure conceptual engagement. By incorporating the work of Kozlowski et al. [10] on cultural dimensions into the CMD workflow, CMD can be used to measure a document's engagement with one or the other pole of binary concepts. When an analyst is interested in measuring how texts engage with a concept that has a culturally obvious opposing concept, using the procedure discussed will allow the analyst to measure whether the texts are engaging more with one pole or the other.

Acknowledgements A replication repository for this paper can be found at: https://github.com/Marshall-Soc/cmd_geometry.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix: Procedures for deriving a semantic directions

Deriving a semantic direction in an embedding space is a specific kind of relation extraction or induction. As such, there are many viable procedures one could use to find the pole of a binary concept in an embedding space. First, the simplest method would involve changing the order of operations used by Kozlowski et al. [10]: average the vectors for the words on each pole and then take the difference between these two averages. Arseniev-Koehler and Foster [1] refer to this method as the "Larsen method" following [13, p. 5]. Kozlowski et al. [10, p. 943 fn8] state that the Larsen method produced "nearly identical results" to theirs.

Second, Arseniev-Koehler and Foster [1] compare the Larsen method to one used in Bolukbasi et al. [3, pp. 42–43], which entails getting the vector offsets of antonym pairs through subtraction, then dividing the resulting vector by the Euclidean norm of the vector offset for those antonym pairs (see also [6]). Arseniev-Koehler and Foster find the results are similar, but the Larsen method was more accurate than this Bolukbasi method.

Third, Bolukbasi et al. [4] offer an additional method involving taking the difference between antonym pairs (they specifically use gendered terms), but then using principal component analysis to find a suitable aggregate from the resulting vector differences.

Finally, for exhaustiveness, there is another procedure which involves measuring individual target words' associations with antonym pairs. This procedure does not, however, define a semantic direction against which any word could be compared and thus cannot be used directly with CMD. Caliskan et al. [5] incorporate this approach into a measure of gender bias in target terms, a technique they refer to as the Word-Embedding Association Test (WEAT). This entails first picking a target term, such as "wrench" or "boat." Then one would take the mean of this target term's distances to female-typed words—such as "girl," "woman," or "lady." Next, one would take the mean of this same term's distances to male-typed words, such as "boy," "man," and "gentleman." Finally, the analyst subtracts the first mean from the second mean, to arrive at a single measure of how strongly associated this target term is to either side of the binary (see also [7]).

References

1. Arseniev-Koehler, A., & Foster, J. (2020). *Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat*. SocArXiv. https://osf.io/preprints/socarxiv/c9yj3/.
2. Atasu, K., Parnell, T., Dünner, C., Sifalakas, M., Pozidis, H., Vasileiadis, V., et al. (2017). Linear-complexity related word mover's distance with GPU acceleration. In J.-Y. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, et al. (Eds.), *2017 IEEE international conference on big data* (pp. 889–896). Boston: IEEE.
3. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Quantifying and reducing stereotypes in word embeddings*. arXiv. https://arxiv.org/abs/1606.06121.
4. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 4349–4357). Curran Associates Inc.
5. Caliskan, A., Bryson, Joanna J, & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.
6. Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). *Understanding undesirable word embedding associations*. arXiv. https://arxiv.org/abs/1908.06361.
7. Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3635–E3644.
8. Goldberg, A. (2011). Mapping shared understandings using relational class analysis: The case of the cultural omnivore reexamined. *American Journal of Sociology*, *116*(5), 1397–1436.
9. Kassambara, A. (2020). ggpubr: 'ggplot2' based publication ready plots. R package version 0.2.5. https://cran.r-project.org/web/packages/ggpubr/ggpubr.pdf. Accessed 11 June 2020.
10. Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905–949.
11. Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In: *International conference on machine learning* (pp. 957–966).
12. Lakoff, George. (2010). *Moral politics: How liberals and conservatives think*. Chicago: University of Chicago Press.
13. Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (pp. 1558–1566). New York: ACM.

14. Makrai, M., Nemeskey, D., & Kornai, A. (2013). Applicative structure in vector space models. In A. Allauzen, H. Larochelle, C. Manning, & R. Socher (Eds.), *Proceedings of the workshop on continuous vector space models and their compositionality* (pp. 59–63). Sofia, Bulgaria: ACL.
15. Mikolov, T, Yih, W.-T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 746–751). aclweb.org.
16. Project Gutenberg. 2020. https://www.gutenberg.org/wiki/Main_Page.
17. Rubner, Y., Tomasi, C., & Guibas L. J. (1998). A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)* (pp. 59–66). IEEE.
18. Sahlgren, Magnus. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33–53.
19. Selivanov, D., Bickel, M., & Wang, Q. (2020) text2vec: Modern text mining framework for R. R package version 0.6. https://cran.r-project.org/web/packages/text2vec/text2vec.pdf. Accessed 11 June 2020.
20. Stoltz, D. S., & Taylor, M. A. (2019). Concept mover's distance: measuring concept engagement via word embeddings in texts. *Journal of Computational Social Science*, *2*(2), 293–313.
21. Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer. **(ISBN 0-387-95457-0)**.
22. Wickham, Hadley. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.
23. Woolley, J. T, & Peters, G. (2008). The American presidency project, Santa Barbara. Available from: http://www.presidency.ucsb.edu/ws.