Political Science Research and Methods (2026), 1–11
doi:10.1017/psrm.2026.10121

REPLICATION

Racialization of the democratic peace in foundational embedding models

Juho Kim

Carnegie Mellon University, USA
Email: juhok@cs.cmu.edu

(Received 20 May 2025 UTC; revised 17 December 2025 UTC; accepted 18 December 2025 UTC)

Abstract

One of the most prominent theories in modern international relations is the democratic peace theory, which posits that democracies tend not to engage in armed conflicts with fellow democracies. Recently, Rathbun, Parker, and Pomeroy examined this effect at a microstructural level, performing surveys, computational studies, and meta-analysis. Of particular originality in their work was embedding analysis, which I review and offer a systematic critique of. Then, I propose a new and improved methodology that addresses its issues while leveraging foundational embedding models. My replication results agree with their findings that the democratic peace effect is racialized, that is, it is associated with racial and ethnocentric motivations. The findings offer a timely guide in conducting embedding analysis in political science research.

Keywords: computational models; international relations; quantitative models; replication; text and content analysis

1. Introduction

The democratic peace theory, which posits that democratic nations are much less likely to wage war against other democracies, is one of the most prominent theories in international relations. While there are controversies and disagreements surrounding its veracity and the mechanism behind it, Tomz and Weeks (2013) found that this expectation also operates at the level of individual citizens, which, in turn, can constrain decision-making by liberal institutions. Recently, Rathbun *et al.* (2024) delved into the motivations behind the public's reluctance, hypothesizing that the concept of democracy is "racialized" and is associated with whiteness, and that this assumption drives the observed peace owing to a shared sense of racial and ethnocentric superiority. To support their hypothesis, the authors performed surveys, computational studies, and meta-analysis to test the degree to which the public associates democracies with whiteness.

The computational study of RPP (Rathbun Parker and Pomeroy, 2024) was unusual in the use of GloVe (Pennington Socher and Manning, 2014). Word vector techniques like word2vec (Mikolov *et al.*, 2013) and GloVe transform words into vectors in a high-dimensional space that represent their semantics. Intuitively, in this vector space, embeddings of semantically similar words are located closer to each other than those that are not. Word embeddings have been widely applied to tasks like visualizing clusters of texts, searching for code snippets that fit a description, solving multiple-choice questions, and selecting similar words (Neelakantan *et al.*, 2022; Müller-Hansen Repke *et al.*, 2023).

The implication of word embeddings in digital humanities and social science is that certain theories, provided they are not driven by a priori principles but instead pertain to human nature, may be

© The Author(s), 2026. Published by Cambridge University Press on behalf of EPS Academic Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

empirically validated by quantifying the meaning of words. The implicit assumption is that the training corpus reflects the human patterns being studied. An obvious example of this includes revealing various types of cultural associations (including biases and stereotypes) in everyday or specialized texts—already extensively explored by prior works (Caliskan *et al.*, 2017; Garg *et al.*, 2018; Kozlowski *et al.*, 2019; Matthews *et al.*, 2022; Sevim *et al.*, 2023)—which also fits the particular use case explored by RPP: Uncovering racial biases in democratic terms and their effect on the democratic peace.

In this replication paper, I scrutinize RPP's embedding analysis and reveal numerous interpretability and statistical issues in their approach. Following up, I propose a new, improved study design in order to address most of the points I bring up. As an additional change, I use third-party foundational embedding models instead of GloVe. The architecture of embedding models, which consists of up to billions of parameters, is based on the designs of large language models (LLMs), which have seen a meteoric rise in relevance and popularity in recent days. Whereas traditional word vector techniques use small contexts and simple neural networks, LLM-based models use *attention* and thus are capable of modeling complex long-range word dependencies. With the advent of embedding models, traditional word vector techniques have become outdated in all practicality, as the newer models are much more powerful, sensitive, and sophisticated. This adds timeliness to my contribution, as RPP relied on the now-superseded GloVe word vectors in their embedding analysis.

My results support RPP's hypothesis, which further reinforces their argument that the democratic public implicitly associates democracies with whiteness, which, in turn, underlies the observed peace effect. In my analysis of the results, I conduct ablation studies, comparing my method to RPP's and that of Tomz and Weeks (2026), who replicated RPP's results with alternative word lists and found contradictory results. In the process, I uncover several insights and pitfalls in the use of embedding analysis, which can benefit researchers interested in incorporating this technique in their toolkit.

2. Embedding analysis of Rathbun, Parker, and Pomeroy (2024)

The core intuition behind RPP's embedding analysis is that, if racialization drives the democratic peace effect, the relationship between democracy and peace should diminish when democracy is associated with non-whiteness, and the opposite should be true for whiteness. The relations between the constructs are quantified using the cosine similarity between their vector estimates, which, in turn, are averages of related word vectors. In an approach reminiscent of the word embedding association test (WEAT) (Caliskan *et al.*, 2017), they defined four word lists to represent the constructs for democracy (**D**), peace (**P**), white (**W**), and non-white ($\overline{\mathbf{W}}$):

$$\begin{aligned}
\mathbf{D} &= \{\text{democracy},\ \text{democratic},\ \text{democratically},\ \text{elect},\ \text{elections},\ \text{elected}\},\\
\mathbf{P} &= \{\text{peace},\ \text{harmony},\ \text{agreement},\ \text{diplomacy}\},\\
\mathbf{W} &= \{\text{white},\ \text{western},\ \text{caucasian},\ \text{european}\},\\
\overline{\mathbf{W}} &= \{\text{non-white},\ \text{non-western},\ \text{non-caucasian},\ \text{non-european}\}.
\end{aligned}$$

RPP defines the "[t]otal similarity" between democratic and peace terms to be

$$s_d = \text{sim}\left(\frac{\sum_{d\in\mathbf{D}} \text{emb}(d)}{|\mathbf{D}|}, \frac{\sum_{p\in\mathbf{P}} \text{emb}(p)}{|\mathbf{P}|}\right),$$

where sim is the similarity function and emb is the embedding function. The "[s]imilarity fixing white" is defined as

$$s_w = \text{sim}\left(\frac{\frac{\sum_{d\in\mathbf{D}} \text{emb}(d)}{|\mathbf{D}|} + \frac{\sum_{w\in\mathbf{W}} \text{emb}(w)}{|\mathbf{W}|}}{2}, \frac{\sum_{p\in\mathbf{P}} \text{emb}(p)}{|\mathbf{P}|}\right).$$

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

Here, whiteness is "fixed" onto the democracy estimate by obtaining the mean of the two estimate vectors. The corresponding value for non-whiteness ($s_{\bar{w}}$) is calculated by using $\overline{\mathbf{W}}$ instead of $\mathbf{W}$.

RPP calculated and plotted the three similarity values s_d, s_w, $s_{\bar{w}}$, along with 90% and 95% confidence intervals. Compared to the baseline similarity between democracy and peace, a small increase in the similarity value (+0.06) was observed when the democracy estimate was fixed with whiteness, whereas a more notable decrease in the similarity value (−0.17) was observed when non-whiteness was fixed, in line with their expectations.

Now, several outstanding issues with RPP's methodology are explored.

2.1. Nonlinearity of the word vector space

Word vector techniques represent a word as a vector in a high-dimensional vector space. Word vectors tend to exhibit locally linear substructures (Pennington Socher and Manning, 2014); however, by the *manifold hypothesis*, "an important perspective on representation learning," complex real-world data is theorized to concentrate near "*strongly nonlinear*" lower-dimensional manifolds in their respective vector spaces (Bengio *et al.*, 2013). Jakubowski *et al.* (2020) even suggest an alternate shape of a "pinched manifold" for word vectors. As a consequence of this global **nonlinearity**, the embedding distribution may twist or curve in space. This property presents a problem in RPP's methodology, namely, in the averaging of word vectors.

In RPP's study, the word vectors were averaged for two reasons. First, in order to generate a vector that estimates a particular construct, they averaged the word vectors of related words. For example, in order to obtain a reasonable estimate for "democracy," they averaged the word vectors of "[d]emocracy terms: democracy, democratic, democratically, elect, elections, elected." In a similar manner, the authors generated such vectors for "peace," "white," and "non-white" constructs. Second, to observe an in/deflation of the similarity between democracy and peace when racialized, they averaged the estimates for "white" and "democracy" to represent "white democracy." The representation of "non-white democracy" was obtained identically. They are not the only ones to take this approach: the same was done by Garg *et al.* (2018) and Kozlowski *et al.* (2019), albeit for different purposes.

This is equivalent to sampling vectors from a distribution and averaging them (Figure 1 visualizes this process). If the manifold on which these vectors lie is linear, then the mean, regardless of what vectors are sampled, is guaranteed to lie on the same manifold by the definition of linearity. However, if the manifold is nonlinear, taking the average risks producing an **out-of-distribution** point. Traditional word vectors are not designed to be averaged for the mean vector to be interpreted as a word that holds some meaning. Indeed, it is unclear how one should interpret the semantics of a point that is located outside the word-vector distribution. One interpretation is that such an out-of-distribution vector has no coherent meaning that pertains to reality, whereas an in-distribution value is more likely to hold tangible semantics. Of course, out-of-distribution points could still be sound enough to allow embedding analysis to be properly conducted.

That said, their methodology can be modified to avoid averaging entirely. Instead of averaging the word vectors and calculating similarities between these estimates, i.e.,

$$s' = \text{sim}\left(\frac{\sum_{x\in\mathbf{X}}\text{emb}(x)}{|\mathbf{X}|}, \frac{\sum_{y\in\mathbf{Y}}\text{emb}(y)}{|\mathbf{Y}|}\right),$$

where $\mathbf{X}, \mathbf{Y}$ are word lists, one can average the similarities from all pairwise combinations of terms related to two different constructs, i.e.,

$$s = \frac{\sum_{(x,y)\in\mathbf{X}\times\mathbf{Y}}\text{sim}(\text{emb}(x), \text{emb}(y))}{|\mathbf{X}||\mathbf{Y}|}. \tag{1}$$

This avoids producing an out-of-distribution embedding, akin to the methodology of Caliskan *et al.* (2017).

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

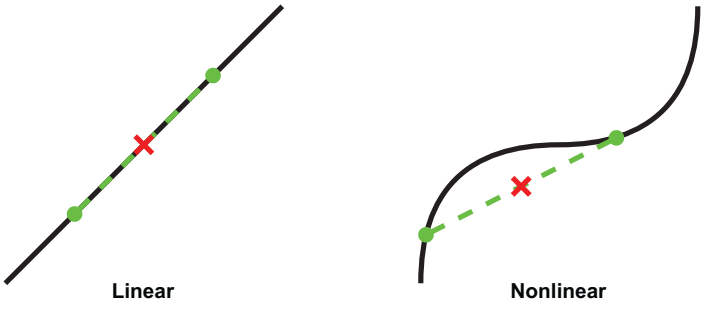


Figure 1. On the left, the mean of the sampled points lies on the original line (note that the line is a 1-dimensional manifold), whereas on the right, it does not.

Additionally, one advantage of embedding models is that **any** text—characters, words, phrases, sentences, paragraphs, and documents—can be projected as an embedding, whereas word vector techniques can only transform a single word to a vector. Using LLM-based embedding models, vector embeddings for a phrase like "white democracy" can be directly obtained instead of having to average estimates for "white" and "democracy." Using the newly proposed methodology, the degree of association between "white democracy" and "peace" can be calculated simply by averaging the pairwise similarities of "white democracy" embeddings and "peace" embeddings. The same can be done to gauge the attenuation of peace tendencies for "non-white democracy."

2.2. Static nature of word vectors

Traditional word vector techniques transform tokens into a **static** point. The term "static" in this context refers to the fact that each word is assigned to a single vector instead of multiple vectors for different contexts. Therefore, a word vector represents an amalgam of meanings over all of the contexts in which the word appears in the training set. For example, the word "white" could refer to white as in the color "white," a "white" lie, or the "white" race. This concern prevented Kozlowski *et al.* (2019) from obtaining a cultural dimension on race, as words "white" and "black" carry a myriad of meanings, including those unrelated to race. To circumvent this, some prior works used common last names for different ethnicities (Caliskan *et al.*, 2017; Garg *et al.*, 2018; Matthews *et al.*, 2022).

RPP handled polysemy largely in two ways. First, they used a corpus of political speeches to train the word vectors instead of a broader multidisciplinary collection of documents. As the vectors were trained on a politically themed dataset, the words' usage would have been mostly limited to political contexts. With this restriction, it is reasonable to expect that the words' vector representations hold a sharper and more focused meaning as it applies in politics. Additionally, RPP created a word list consisting of similar words and averaged their word vectors. Ignoring the interpretability issues with averaging, it is reasonable to expect the mean vector to contain a strong signal for the shared definition.

However, Tomz and Weeks (2026) found RPP's word lists to be problematic, characterizing them as being "unofficial," "ad hoc," and unjustified. Instead, they proposed alternative word lists based on the United States government's federal reporting guidelines,[1] and replicated their analysis, where the new word lists led to results that contradict RPP. In this paper, I perform observational studies using both sets of word lists to resolve this tension.

[1]See: https://www.federalregister.gov/d/2024-06469/p-177.

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

Additionally, as mentioned, RPP calculated an estimate of "(non-)white democracy" by averaging the estimates for "(non-)white" and "democracy." This may not give rise to an acceptable representation of "(non-)white democracy" since it is possible that the "white" estimate is more similar to "peace" than the "non-white" estimate is to begin with, and this difference simply carries over even when they are averaged with the "democracy" estimate. If true, this possibility has the potential to invalidate the RPP results. In my ablation studies, I find weak evidence showing that this is the case.

The use of LLM-based embedding models provides a simple and convincing solution to all of these issues. The embedding for the phrase "(non-)white democracy" is a static vector that represents the entire phrase (not just the mean of each word in isolation). Also, instead of averaging the embeddings of terms related to "(non-)white democracy," by averaging the similarities of all pairs of "(non-)white democracy" terms and "peace" terms, as suggested in my alternate methodology, the ambiguity introduced by averaging different words can be circumvented.

2.3. Bootstrapped confidence intervals

In obtaining the confidence intervals of the results, RPP performed evaluations over a thousand dictionary permutations, as done by Garg *et al.* (2018). Vallebueno *et al.* (2024) criticize that this approach is "computationally intractable on large datasets" and only "address[es] uncertainty from document or lexicon selection, even though embeddings are parameters of a data-generating process on word co-occurrences, not on collections of documents or sets of words." Aside from their criticisms, the number of RPP's terms for each construct of interest ranges from 4 to 6, which is too small to produce a sufficiently rich set of bootstrap samples. The word lists proposed by Tomz and Weeks (2026) do not add new democratic and peace terms, so the argument about the insufficient richness of bootstrap samples still holds. The same consideration also prevents one from conducting WEAT-style tests. (Caliskan *et al.*, 2017)

One way to address this is to use GloVe-V (Vallebueno *et al.*, 2024), which, aiming to directly address this challenge, integrates statistical uncertainty into the model, allowing error variance estimation and hence sounder confidence interval calculation. Unfortunately, the use of different LLM-based embedding models in this replication does not address this issue. That said, demonstrating that the same directional shifts occur consistently across most of the tested models would still serve as a strong empirical evidence in support of the hypothesis.

3. New methodology

In our observational study, the same base "white," "non-white," "democracy," and "peace" terms used by RPP are used. However, "white" and "non-white" terms are never directly converted into embeddings, but instead are combined with the "democracy" terms to fix white or non-white information prior to encoding. The sets of terms considered in this analysis are shown below:

$$\mathbf{D} = \{\text{democracy},\ \text{democratic},\ \text{democratically},\ \text{elect},\ \text{elections},\ \text{elected}\},$$
$$\mathbf{P} = \{\text{peace},\ \text{harmony},\ \text{agreement},\ \text{diplomacy}\},$$
$$\mathbf{W} = \{\text{white},\ \text{western},\ \text{caucasian},\ \text{european}\} \times \mathbf{D},$$
$$\overline{\mathbf{W}} = \{\text{non-white},\ \text{non-western},\ \text{non-caucasian},\ \text{non-european}\} \times \mathbf{D}.$$

Here, **D**, **P**, **W**, and $\overline{\mathbf{W}}$ are the set of democracy, peace, white-democracy, and non-white-democracy terms, respectively. Note that **W** and $\overline{\mathbf{W}}$ are the Cartesian products of (non-)white terms and democracy terms. For example, "western democracy" is a member of **W**. I also repeated the analysis using the alternative set of word lists[2] from Tomz and Weeks (2026):

$$\mathbf{W} = \{\text{white}, \text{english}, \ldots, \text{scottish}\} \times \mathbf{D}, \qquad \overline{\mathbf{W}} = \{\text{asian}, \text{chinese}, \ldots, \text{mayan}\} \times \mathbf{D}.$$

[2]The full lists are in the supplementary code.

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

Table 1. Average semantic similarities between peace and (non-)white democratic terms (curated by RPP) in foundational embedding models

Model	No race	White	Non-white
`text-embedding-004` (Google)	0.360	0.325	**0.304**
`embedding-001` (Google)	0.628	0.577	**0.554**
`text-embedding-3-small` (OpenAI)	0.298	0.208	**0.183**
`text-embedding-3-large` (OpenAI)	0.261	0.196	**0.177**
`text-embedding-ada-002` (OpenAI)	0.800	0.770	**0.764**
`all-mpnet-base-v2`	0.245	0.206	**0.144**
`multi-qa-mpnet-base-dot-v1`	17.091	15.988	**14.449**
`all-distilroberta-v1`	0.235	0.195	**0.155**
`all-MiniLM-L12-v2`	0.310	0.229	**0.177**
`multi-qa-distilbert-cos-v1`	0.195	0.110	**0.100**
`all-MiniLM-L6-v2`	0.313	0.203	**0.149**
`multi-qa-MiniLM-L6-cos-v1`	0.189	0.136	**0.119**
`paraphrase-multilingual-mpnet-base-v2`	0.284	0.286	**0.263**
`paraphrase-albert-small-v2`	0.286	**0.212**	0.216
`paraphrase-multilingual-MiniLM-L12-v2`	0.350	0.255	**0.242**
`paraphrase-MiniLM-L3-v2`	0.283	**0.179**	0.182
`distiluse-base-multilingual-cased-v1`	0.421	0.273	**0.233**
`distiluse-base-multilingual-cased-v2`	0.387	0.254	**0.217**

Note: The minimum value for each model is bolded.

The average semantic similarity between democracy terms and peace terms is

$$s_d = \frac{\sum_{(d,p)\in \mathbf{D}\times\mathbf{P}} \text{sim}(\text{emb}(d), \text{emb}(p))}{|\mathbf{D}||\mathbf{P}|}.$$

Here, all pairwise similarity values are calculated, which are then averaged to yield a single average semantic similarity value. The average similarity between white-democracy and peace terms (s_w) and that between non-white-democracy and peace terms ($s_{\bar{w}}$) is calculated identically except that $\mathbf{W}$ and $\overline{\mathbf{W}}$ are used, respectively, instead of $\mathbf{D}$. Assuming that the similarity values are independent (which may not be safe), I also conducted the one-sided Welch's t-test for the means of cosine similarities between white-democracy and peace embeddings (s_w) and their non-white counterparts ($s_{\bar{w}}$), and these results are relegated to the supplementary.

Embedding models are often tailored to different similarity functions, which can have different score scales. If they were to be plotted together, there is a need to "normalize" each value so they are of roughly equal magnitude. Before plotting, each average similarity was normalized as

$$s_w^* = \frac{s_w - s_d}{s_d}.$$

The corresponding value for non-white democracy $s_{\bar{w}}^*$ is calculated by replacing s_w with $s_{\bar{w}}$.

In the selection of third-party foundational embedding models, I aimed to curate a balanced mix of open-source and proprietary options. I selected all models offered by Google Gemini and OpenAI, as well as every recommended model from the SentenceTransformers library. The complete list of models is available in Table 1.

Foundational embedding models used in this analysis are typically fitted from a general-purpose dataset across many subjects in both formal and informal contexts, with some in multilingual settings. This approach differs from the one RPP took, where a specialized corpus of English political speeches was used instead of pre-trained GloVe embeddings. Past works investigating the use of embedding analysis in political settings found that it is not necessary to specialize a model to yield sound results. Indeed, Rodriguez and Spirling (2022) found that pre-trained, unspecialized models frequently match the performance of or even outperform specialized models in several politically

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

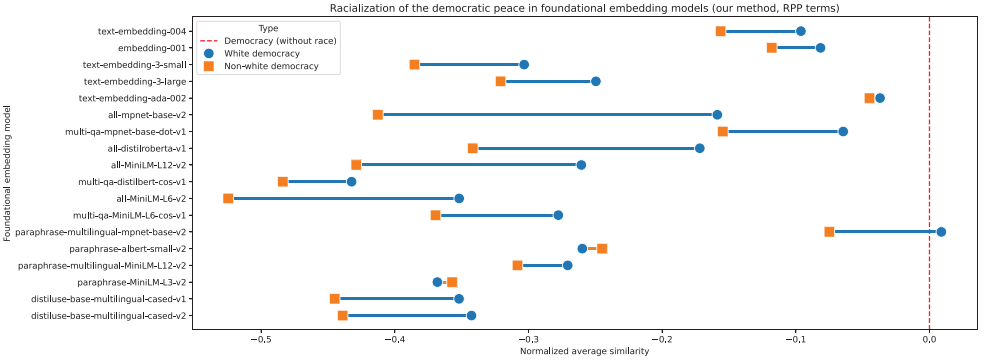


Figure 2. Normalized average semantic similarities between peace and white and non-white democratic terms, curated by RPP.

themed case studies. The fact that the goal of this study is to quantify the general public's perception of democracies further justifies the choice of using pre-trained foundational models—certainly trained on everyday conversations during pre-training—instead of transfer learning.

4. New results

The average semantic similarities between peace and (non-)white democratic terms given by RPP are tabulated in Table 1, and their normalized values are plotted in Figure 2. In the plot, blue circles represent the normalized average semantic similarities between white-democracy and peace terms, while orange squares represent their non-white counterparts. The lines between them are colored blue for models in which white-democracy terms are more semantically similar to peace terms than non-white-democracy terms are. Otherwise, the lines are colored orange. The dotted red vertical line represents the base semantic similarity of peace and democracy terms. If a point lies to the left of the line, it means that the corresponding terms are, on average, less similar to peace terms than democracy terms are, and vice versa.

Sixteen out of 18 models exhibited the directional trend expected of the racialization of democratic peace, and the differences ranged from noticeably large to marginal. I repeated the study using the alternative set of word lists from Tomz and Weeks (2026), the normalized results of which are plotted in Figure 3. Here, 14 out of the 18 models exhibited the directional trend expected of the racialization of democratic peace, although the differences between white and non-white democratic terms were less than before.

To summarize, using my methodology, a general directional trend that strongly supports RPP's hypothesis is observed regardless of which word lists are used. Thus, my results agree with RPP but contradict Tomz and Weeks (2026). This introduces a new tension between my results and those of Tomz and Weeks, which I seek to resolve by ablating on my methodology.

4.1. Is averaging word embeddings wrong?

The two differences between my methodology and that of RPP are that a) I did not take the average of embeddings and b) I used embedding models instead of GloVe. To identify if taking the average gave rise to the aforesaid tension, I replicated RPP's methodology with both sets of word lists while using foundational embedding models.

The normalized results with the original word lists and methodology from RPP are plotted in Figure 4. While the general directional trend is preserved, it is somewhat surprising to observe

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

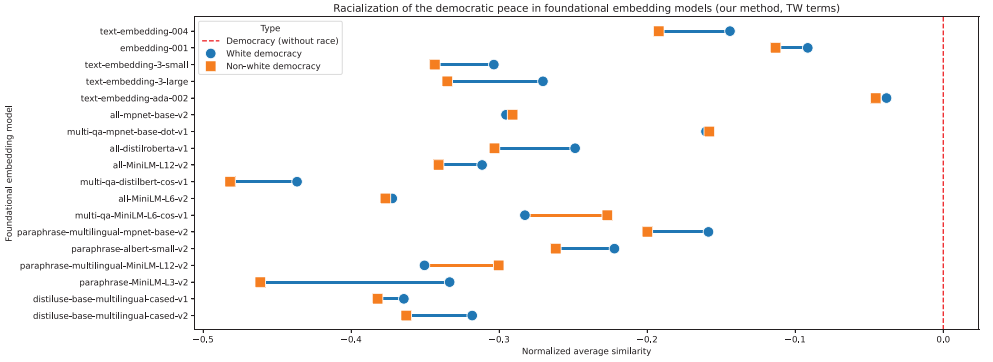


Figure 3. Normalized average semantic similarities between peace and white and non-white democratic terms, curated by Tomz and Weeks (2026).

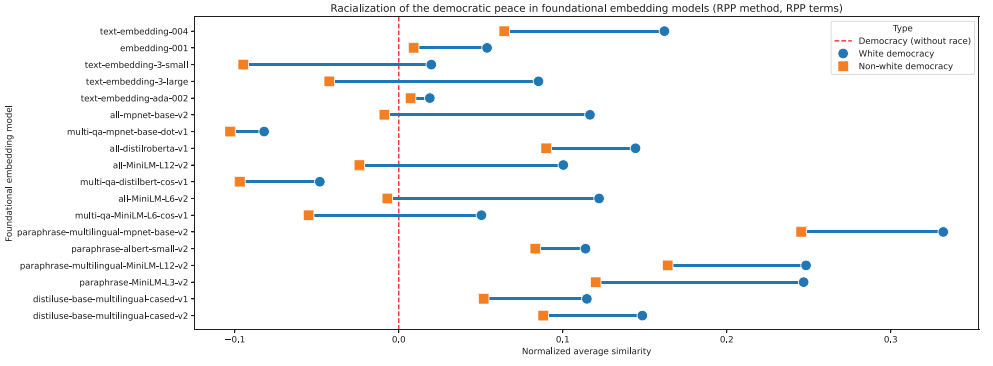


Figure 4. Normalized semantic similarities between mean peace and white and non-white democratic embeddings, following the established method of RPP and their word lists.

the general horizontal shift in the normalized similarity values. This suggests that interpreting the **absolute** similarity values when performing embedding analysis carries risks of making overreaching interpretations, as small changes in the methodology can give rise to wide shifts in the values. However, the offsets between the white and non-white scores are well-preserved despite this shift, which suggests that, when performing embedding analysis, it is much more advisable to make interpretations by comparing similarities **relative** to others.

Similarly, the results using the word lists from Tomz and Weeks (2026) and the methodology of RPP are plotted in Figure 5. The new plot showcases a result that is remarkably different from those in the previous plots: Whereas a clear directional trend was observable in all three previous plots, the directional trend in Figure 5 is mixed. Clearly, when the alternative word lists and RPP's original methodology are used, the result does not seem to support RPP's hypothesis, which aligns with the finding of Tomz and Weeks!

Thus, the source of the tension between my finding and Tomz and Weeks seems to be in the averaging of word embeddings, not in the use of foundational embedding models. If I establish that RPP's original methodology is flawed, I would be able to resolve the tension between my result and that of Tomz and Weeks. In the next subsection, I establish this condition by demonstrating weak evidence that averaging word embeddings is an unsound practice.

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

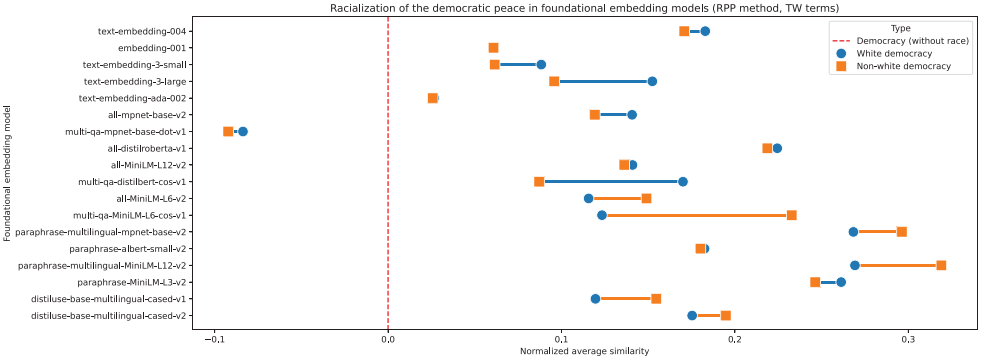


Figure 5. Normalized semantic similarities between mean peace and white and non-white democratic embeddings, following the established method of RPP, while using the word lists from Tomz and Weeks (2026).

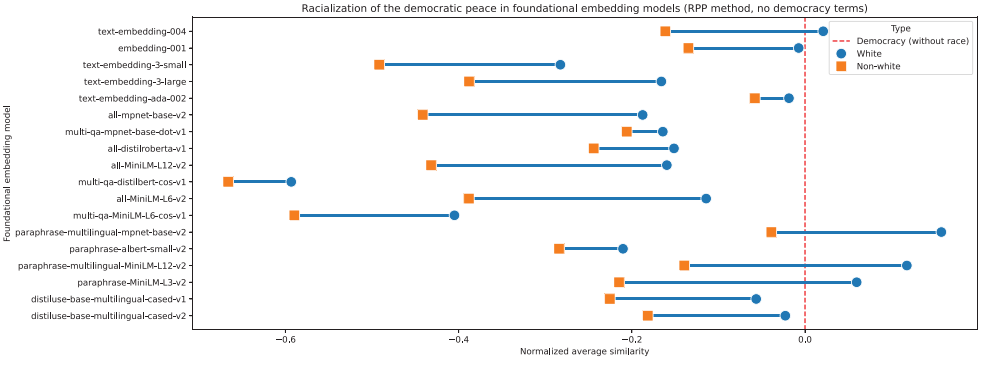


Figure 6. Normalized semantic similarities between mean peace and white and non-white embeddings (no democracy terms involved), following RPP's established method.

4.2. Ignoring democratic terms

In order to gain more insight, I performed another ablation study, following RPP's methodology, while ignoring the democracy estimate entirely. I simply calculated the similarity score between the peace estimate and the (non-)white estimates. The new values are plotted in Figure 6.

The new plot reveals a novel insight into the orientation of the estimates. The white and non-white estimates are, in general, no more similar to the peace estimate than the democracy estimate is. However, Figure 4 shows that, when the mean is taken between the (non-)white estimate and the democracy estimate, it yields a new vector that is much more similar to the peace estimate than both of its constituent vectors. This reveals that the vectors for (non-)white, democracy, and peace estimates are roughly **coplanar** and the peace estimate is an approximate **angle bisector** between the other two.

The new results also confirm one of the previous concerns: The white estimate is more similar to the peace estimate than the non-white estimate is, and this feature may have propagated to the estimate for "(non-)white democracy," thus unduly influencing the RPP result. The results from this ablation study serve as weak evidence that generating an estimate by taking the average of word embeddings is flawed.

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

5. Discussion

When interpreting results from embedding analysis, there is a risk of making overreaching interpretations. As these models are trained on real-world data, which contains existing biases and stereotypes, the output word vectors may also encode such prejudice. This is especially true for multilingual models, which may encode stereotypes from different cultures that can leak to English word vectors, owing to the remarkable ability of LLMs to generalize. AI firms offering third-party embedding models often lack transparency about their training set, further making it challenging to judge the potential biases of their outputs. It is therefore important to be mindful of this shortcoming and acknowledge the black-box nature of embedding models.

Additionally, it is important to be careful not to invoke "self-fulfilling prophecies." Certain theories may be extremely well-known to the extent that there are extensive discussions about them. When trained on data containing these discussions, the resulting embeddings may encode information about the hypothesis as latent features, which can give the impression that the pattern is truly reflected in the data.

6. Conclusions

In this paper, I analyzed the embedding analysis of Rathbun *et al.* (2024) as part of their investigation on whether ethnocentrism underpins the peace effect in the democratic peace theory. I offered a systematic critique of their methodology and proposed a new modernized methodology addressing some of its issues. Overall, my replication supports RPP's hypothesis on the racialization of the democratic peace and uncovers novel insights about the theory and practice of embedding analysis. It remains to be seen what other theories in political science can be supported by new computational tools like embedding analysis.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2026.10121.

Funding statement. The author has no relevant funding statement to declare.

Competing interests. The author has no relevant competing interests to declare.

References

Andrea V, Handan-Nader C, Manning CD and Daniel EH (2024) Statistical uncertainty in word embeddings: GloVe-V. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Miami, Florida, USA: Association for Computational Linguistics, pp. 9032–9047.

Aylin C, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* **356**(6334), 183–186.

Bengio Y, Courville A and Vincent P (2013) Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8), 1798–1828.

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

Jakubowski A, Gasic M and Zibrowius M (2020) Topology of word embeddings: Singularities reflect polysemy. In *Proceedings of the Joint Conference on Lexical and Computational Semantics (*SEM)*. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 103–113.

Kozlowski AC, Taddy M and Evans JA (2019) The geometry of culture: analyzing the meanings of class through word embeddings. *American Sociological Review* **84**(5), 905–949.

Müller-Hansen F, Repke T, Baum CM, Brutschin E, Callaghan MW, Debnath R, Lamb WF,Low S, Lück S, Roberts C, Sovacool BK and Minx JC (2023) Attention, sentiments and emotions towards emerging climate technologies on Twitter. *Global Environmental Change* **83**, 102765.

Matthews S, Hudzina J and Sepehr D (2022) Gender and racial stereotype detection in legal opinion word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Vancouver, Canada: Association for the Advancement of Artificial Intelligence, pp. 12026–12033.

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

Mikolov T, Chen K, Corrado G and Dean J (2013) Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations Workshop (ICLRW)*. Scottsdale, Arizona.

Neelakantan A, Tao X, Puri R, Radford A, Han JM, Tworek J, Yuan Q, Tezak N, Kim JW, Hallacy C, Heidecke J, Shyam P, Power B, Nekoul TE, Sastry G, Krueger G, Schnurr D, Such FP, Hsu K, Thompson M, Khan T, Sherbakov T, Jang J, Welinder P and Weng L (2022) Text and code embeddings by contrastive pre-training.

Pennington J, Socher R and Manning C (2014) GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.

Rathbun BC, Parker CS and Pomeroy C (2024) Separate but unequal: Ethnocentrism and racialization explain the "Democratic" peace in public opinion. *American Political Science Review* **119**(2), 621–636.

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

Sevim N, Sahinuc F and Koc A (2023) Gender bias in legal corpora and debiasing it. *Natural Language Engineering* **29**(2), 449–482.

Tomz M and Weeks J (2013) Public opinion and the democratic peace. *American Political Science Review* **107**(4), 849–865.

Tomz M and Weeks J (2026) Race, democracy, and public support for war. *American Political Science Review* **120**(2), 707–724.

https://doi.org/10.1017/psrm.2026.10121 Published online by Cambridge University Press

Cite this article: Kim J (2026) Racialization of the democratic peace in foundational embedding models. *Political Science Research and Methods*, 1–11. https://doi.org/10.1017/psrm.2026.10121