Political Analysis (2025), **00**, 1–21
doi:10.1017/pan.2025.6

ARTICLE

Constructing Vec-tionaries to Extract Message Features from Texts: A Case Study of Moral Content‡

Zening Duan[1], Anqi Shao[2], Yicheng Hu[3], Heysung Lee[1], Xining Liao[1], Yoo Ji Suh[1], Jisoo Kim[1], Kai-Cheng Yang[4], Kaiping Chen[2] and Sijia Yang[1]

[1]School of Journalism and Mass Communication, University of Wisconsin–Madison, WI, USA; [2]Department of Life Sciences Communication, University of Wisconsin–Madison, WI, USA; [3]Department of Chemical and Biological Engineering, University of Wisconsin–Madison, WI, USA; [4]Network Science Institute, Northeastern University, Boston, MA, USA

Corresponding author: Sijia Yang; Email: sijia.yang@alumni.upenn.edu

(Received 26 February 2024; revised 10 May 2024; accepted 4 June 2024)

Abstract

While researchers often study message features like moral content in text, such as party manifestos and social media posts, their quantification remains a challenge. Conventional human coding struggles with scalability and intercoder reliability. While dictionary-based methods are cost-effective and computationally efficient, they often lack contextual sensitivity and are limited by the vocabularies developed for the original applications. In this paper, we present an approach to construct "vec-tionaries" that boost validated dictionaries with word embeddings through nonlinear optimization. By harnessing semantic relationships encoded by embeddings, vec-tionaries improve the measurement of message features from text, especially those in short format, by expanding the applicability of original vocabularies to other contexts. Importantly, a vec-tionary can produce additional metrics to capture the valence and ambivalence of a message feature beyond its strength in texts. Using moral content in tweets as a case study, we illustrate the steps to construct the moral foundations vec-tionary, showcasing its ability to process texts missed by conventional dictionaries and to produce measurements better aligned with crowdsourced human assessments. Furthermore, additional metrics from the vec-tionary unveiled unique insights that facilitated predicting downstream outcomes such as message retransmission.

Keywords: computational text analysis; message feature; moral content; word embedding; optimization; crowd-sourcing

Edited by: Jeff Gill

1. Introduction

Social scientists from various disciplines have worked on improving the quantitative measurement of message features, such as emotions (Brady *et al.* 2017), uncivil and gendered language (Chen, Duan, and Kim 2024; Theocharis *et al.* 2016), and more recently, moral intuitions (Clifford and Jerit 2013; Graham *et al.* 2013; Weber *et al.* 2021; Zhou *et al.* 2022). This exploration extends across diverse text sources, including government records, newspapers, social media posts, and other unstructured textual repositories. However, quantifying message features from texts presents a formidable challenge. For example, human coding cannot easily scale up to process "big data" (Hopkins and King 2010), or in some cases, is suboptimal to alternative measurement strategies such as crowdsourcing, particularly when intercoder reliabilities fall short of conventional thresholds (Weber *et al.* 2021). The rise of computational content analysis methods, notably text-as-data approaches (Grimmer, Roberts, and Stewart 2022), has popularized the use of dictionaries as a low-cost, quick-to-use measurement strategy

‡This article has been updated since its original publication. A notice detailing these changes can be found here: https://doi.org/10.1017/pan.2025.9

for handling large-scale textual data. However, this approach has inherent limitations, lacking sensitivity to context-specific applications and often encountering difficulties in extracting signals from short-format texts like tweets due to its fixed and limited vocabulary.

This study introduces "vec-tionaries," a novel computational method for extracting message features. We use moral content as a case study to demonstrate its advantages. The Moral Foundations Theory (MFT) (Graham, Haidt, and Nosek 2009; Haidt 2012) suggests that individuals' moral intuitions are rooted in six major psychological systems or foundations, including Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, and Liberty/Oppression. Each of these foundations acts like a "taste bud," allowing individuals to quickly judge situations in the social world that uphold or violate these foundations through gut-like reactions of likes and dislikes (Haidt 2012). For instance, the Care/Harm foundation involves sensitivity toward the suffering of vulnerable beings, such as war refugees, while those attuned to Authority/Subversion tend to prioritize social hierarchy and tradition. MFT has reshaped scholarly understanding of morality and how it relates to the formation of political attitudes, expressions, and behaviors. A growing body of research demonstrates that moral foundations play a crucial role in fueling partisan disagreements on environmental attitudes (Feinberg and Willer 2013), candidate trait evaluations (Clifford 2014), and voting choices (Jung 2020).

Although initially developed as a psychological theory, MFT appeals broadly to social scientists interested in studying morality-related content in various types of messages (e.g., news coverage, social media posts, entertainment media), who often treat moral rhetoric and moral appeals as a category of latent message features that invoke and appeal to moral foundations. In political science, motivated by the belief that moral words may do "the work of politics," scholars have used quantitative measurement advancements to bridge the gap between MFT—a theory concerned with the formation of pluralistic moral foundations in individuals' minds and moral judgment—and research focused on systematically analyzing morality-related content across different political contexts, including party manifestos (Jung 2020), speeches (Graham *et al.* 2009), and state legislatures (Mucciaroni 2011). Methodological advancements in measuring moral content have helped expand the applications of MFT to outcomes with broader societal impacts, such as online information diffusion (Brady *et al.* 2017), gender stereotype (Chen *et al.* 2024), hate speech (Solovev and Pröllochs 2023), political participation (Jung 2020), persuasion (Kaplan *et al.* 2023; Yang and Yang 2023), and public opinions on sociopolitical controversies (Clifford and Jerit 2013; Feinberg and Willer 2013). However, measuring moral content as a latent message feature presents significant methodological challenges, such as difficulty achieving inter-coder reliability in conventional content analytical approaches relying on a small number of human annotators (Weber *et al.* 2021; Hopp *et al.* 2021). This has motivated scholars to explore new approaches like crowdsourcing and machine learning (Hoover *et al.* 2020; Hopp *et al.* 2021). Our vec-tionary approach aims to address these conceptual and methodological challenges, providing an accessible, interpretable, and scalable tool for extracting moral content from textual data.

Our vec-tionary approach leverages the semantic relations between validated dictionary words encoded in pre-trained word embeddings, where the message features can be represented as semantic axes residing in the same semantic vector space (An, Kwak, and Ahn 2018; Kozlowski, Taddy, and Evans 2019). Our model then identifies these axes through a nonlinear optimization algorithm. Users can then project unseen messages onto these axes to measure the message features of interest. Compared with the dictionary approach, which only contains the semantic information of a limited vocabulary, a vec-tionary incorporates additional signals from other words outside the original dictionary's vocabulary by exploiting their embeddings-based semantic relations. Moreover, pre-trained word embeddings allow a vec-tionary to capture contextual information in documents and quantify additional properties of the message feature such as *Valence* and *Ambivalence*, without relying on human-labeled data for supervised classifier training. While our study focuses on moral content to illustrate the measurement advantages, conceptual foundations, and implementation protocols of vec-tionaries, we note that the methodology for constructing vec-tionaries extends beyond moral content and can be applied to measure various message features, such as emotions, frames, incivility, and many more.

Next, we overview the strengths and weaknesses of existing computational methods for measuring moral content in Section 2. Section 3 introduces our vec-tionary approach and three metrics derived from it to capture different aspects of moral content in texts. Section 4 compares vec-tionaries to the state-of-art moral foundations dictionary (MFD) crowdsourced annotations from two million COVID-19 tweets, showing that our approach is superior to, or at least on par with, existing methods for measuring moral content. Section 5 applies our moral foundations vec-tionary to study extracting moral content from the same tweet corpus, predicting retweets, and demonstrating additional value in enhancing empirical research on moral content. Section 6 concludes; proofs, illustrations, and supporting information are in Sections A–L of the Supplementary Material.

2. Existing Computational Methods to Measure Moral Content

Dictionaries and word embeddings are two of the most prominent methods to extract moral content from textual data. In this section, we provide an overview of these two measurement strategies and discuss their strengths and limitations.

2.1. Moral Foundations Dictionaries

In early work, Graham et al. developed the first MFD using frequencies of foundation-relevant words (Graham *et al.* 2009), particularly synonyms and antonyms, to measure differences in moral values between liberal and conservative sermons. However, the original MFD had fewer words (on average, 32 for each moral foundation) than many other dictionaries. Frimer and colleagues introduced the MFD 2.0, a more sophisticated version of the first MFD (Frimer *et al.* 2017), by proposing a much larger set of candidate words. Subsequently, the extended MFD (eMFD) further expanded the list to encompass approximately 3,270 English words associated with five moral foundations with varying weights (Hopp *et al.* 2021). Deviating from its ancestors, eMFD assigns each word to all five moral foundations instead of exclusively to a single moral foundation. Additionally, eMFD is constructed from text annotations generated by a group of human coders ($n = 557$) rather than a few trained coders. As the latest addition to MFT, the Liberty/Oppression foundation was absent in most existing dictionaries, including those mentioned above. To address this, Araque et al. introduced LibertyMFD, a foundation-specific lexicon to operationalize this moral foundation (Araque, Gatti, and Kalimeri 2022).

Word-count-based method has made significant strides in the textual analysis of moral content (Solovev and Pröllochs 2023), especially excels at interpretability. By employing pre-established word lists, this method provides direct insights into the contributing words that define the message feature. Nevertheless, this approach has some drawbacks. Its effectiveness largely depends on the vocabulary included in the dictionary, any omission of a word results in reduced coverage. Moreover, this method often overlooks the context in which words appear. A single word might bear different meanings based on its surrounding context, a nuance often missed, making it difficult to generalize a dictionary developed in one specific context to others. The dictionary approach often suffers from inflexibility, particularly when adapting or extending the dictionary to accommodate evolving linguistic nuances, a task that can be labor-intensive. All present notable challenges and call for improvement. As a response, Garten et al. introduced the Distributed Dictionary Representation (DDR) approach (Garten *et al.* 2018) and An et al. proposed the SEMAXIS framework, both utilizing word-embedding to better quantify short-form texts from contextually dependent data (An *et al.* 2018). In the next section, we provide detailed explanations of these word embedding approaches and then illustrate how our moral foundations vec-tionary is designed and implemented building upon these efforts.

2.2. Word Embeddings and DDRs

In the field of natural language processing, significant progress has been made in learning effective representations of words as vectors in high-dimensional semantic spaces (Mikolov, Yih, and Zweig 2013). These vectors, known as word embeddings, have been applied to analyze embedded semantic meanings of concepts such as equality (Rodman 2020), class (Kozlowski *et al.* 2019), and incivility (Liang, Ng, and Tsang 2023) across spatial, temporal, and cultural contexts.

In a word embedding model, each unique word appearing in a document is represented by a vector (Mikolov *et al.* 2013; Pennington, Socher, and Manning 2014) that positions it in a high-dimensional geometric space in relation to every other unique word. A word's adjacent neighbors in the vector space are usually words with related meanings, including the word's own syntactic variants or synonyms. The geometric relationship, or distance, between two vectors, signals the semantic similarity or divergence of the corresponding words. Such distance, or the lack thereof, is commonly quantified by the cosine similarity between these two vectors. Many word embedding methods have been proposed in the past decade. Among these, Word2vec stands out as one of the most widely used. Introduced by Mikolov and colleagues in 2013, Word2vec employs a two-layer neural network to process text by vectorizing words: its input is a text corpus, and the output is a set of vectors that represent words in that corpus. In our following analyses, we demonstrate how even a plain word embedding model can be integrated with a dictionary to enhance the model's performance in measuring latent moral signals from texts.

Words that are geometrically clustered can indicate a latent semantic concept, constructing a representation of a latent concept is thus analogous to building a word representation in the vector space. The DDR approach (Garten *et al.* 2018) utilizes the average of vector representations of the words in a dictionary to represent a given concept or an embedded message feature like moral content in our context. For instance, the *care*-relevant content can be represented by computing the average of vectors associated with *care*-related words like [*kindness*, *compassion*, *nurture*, *empathy*]. The DDR approach, then, facilitates the computing of a continuous similarity metric between a moral foundation and a text. This is achieved by projecting the text into the same vector space and then calculating the similarity between the vectorized text and the averaged dictionary word vectors representing the moral foundation.

Concepts such as moral foundations often entail valences, such as *care* and *harm* serving as the two anchors for the *Care/Harm* moral foundation. Using the DDR approach, one can construct a concept representation of, for example, the virtue of *care* by averaging all relevant words associated with care per se. However, the representation of the vice of *harm* remains a challenge—even though one can similarly construct it by averaging the vectors of *harm*-related words, this axis usually would not be geometrically positioned as the opposite anchor to care on the same *Care/Harm* axis. To address this limitation, An and colleagues proposed SEMAXIS (An *et al.* 2018), a framework that creates an integrated vector axis for a target concept, encompassing both its positive and negative aspects (e.g., virtue and vice for a moral foundation) as the two opposing anchors on the shared axis, also see a similar approach (Sagi and Dehghani 2014). It can be understood as a "concept axis" in a vector space. Analyzing such an axis allows us to measure the semantic similarity of documents composed of individual words relative to these concept axes.

In this context, a concept axis, or a moral axis in our case, is anchored by an antonym pair, such as *Care–Harm*, *Fairness–Cheating*, or *Authority–Subversion*. Each antonym pair typically includes a set of the most positive (or rightness) words on one end and the most negative (or violation) words on the other (An *et al.* 2018). To calculate the concept axis, for example, the *Care/Harm* moral axis, the positive anchor (i.e., the virtue of *care*) is first built using the DDR method by averaging the vectors of all positive words, and a similar process is applied for building the negative anchor (i.e., the vice of *Harm*). SEMAXIS then finds the semantic axis that connects the negative anchor with the positive by taking the difference between the averaged vectors of two sets of pole words, i.e., the positive and negative words, related to this moral foundation (An *et al.* 2018). Thus, once the moral axis vector is obtained, researchers can compute the cosine similarity between a word vector and the axis to quantify the moral relevance of a single word or a text (Kwak *et al.* 2021). That being said, integrating all pole words from a well-established dictionary into building moral axes comes with several challenges awaiting solutions: words often contribute differently to a specific concept they are associated with, for instance, the word "*murder*" likely contributes more to the *Care/Harm* axis than "*slap*," and assigning the right weight to each pole word when constructing the concept axis is both conceptually and statistically challenging. In the following sections, we will provide a detailed explanation of how our model has effectively tackled these challenges using an optimization algorithm and thus lifting the advantages of two conventional approaches into one.

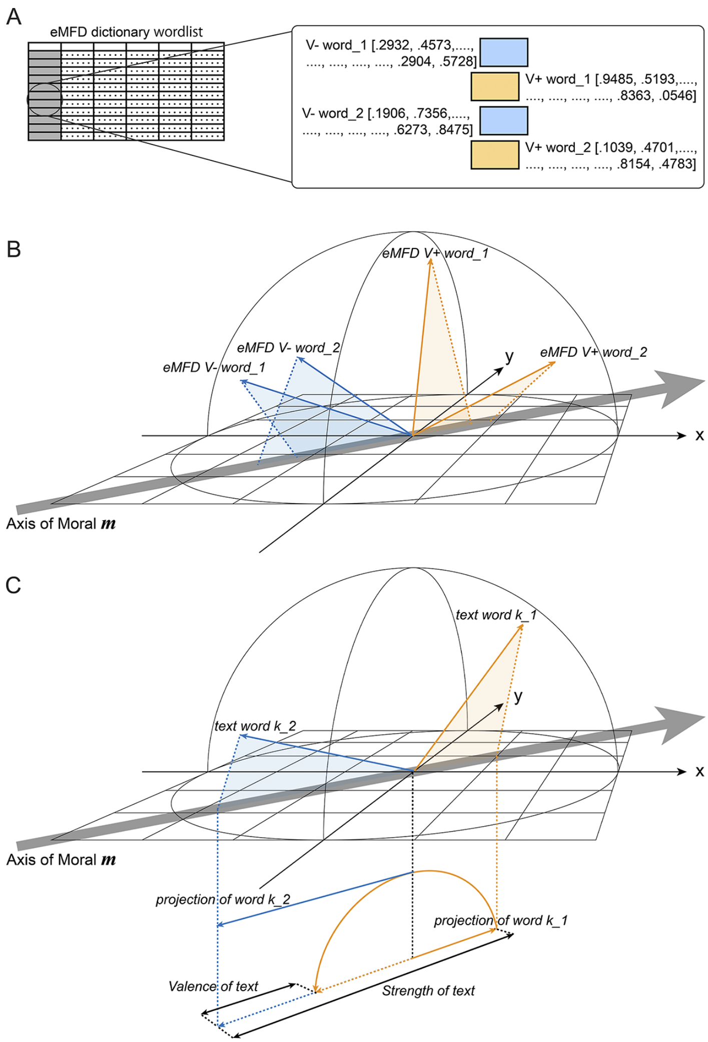

Figure 1. The model pipeline of moral foundations vec-tionary.

3. The Vec-tionaries Approach and the Construction of the Moral Foundations Vec-tionary

In this work, we introduce a novel framework called *vec-tionary* that integrates the well-established dictionary (i.e., eMFD) with word embedding models to measure moral content embedded in textual data. Specifically, we constructed the proposed moral foundations vec-tionary following three steps for a chosen moral foundation: (1) vectorizing words in the eMFD based on a preselected word embedding model, (2) estimating the axis for the targeted moral foundation through a nonlinear optimization algorithm, and (3) calculating the geometric distance between an unseen text and this estimated moral axis in the same vector space to derive metrics of interest such as *Strength*, *Valence*, and *Ambivalence* of a targeted moral foundation. See Figure 1 for an illustration of the pipeline to construct the moral foundations vec-tionary.

We assume that the axes representing different moral foundations exist in the shared vector space with words contained in the eMFD, and our goal is to uncover these axes' geometric coordinates. We leverage eMFD's large vocabulary and crowdsourced "weights indicating the semantic relationships with each moral foundation. We treat each weighted word as an "observed signal" of the latent moral axis.

https://doi.org/10.1017/pan.2025.6 Published online by Cambridge University Press

Employing a nonlinear optimization algorithm, we iteratively update our estimates for the coordinates of the moral axes to best account for the observed weighted words from eMFD, which are themselves embedded in the same vector space. In our analyses, we used the 300-dimensional embeddings from the word2vec model, which covers nearly 3 million words and phrases. In the following sections, we present the technical details of the vec-tionary approach.

3.1. Mathematical Framework

As illustrated in Figure 1, first, we transformed each eMFD word to word vectors in the semantic space. According to the assumption of the eMFD, each word i is linked to all five moral foundations (except for Liberty/Oppression, a recently added moral foundation not included in the eMFD), albeit with varying weights. The analytical goal of the moral foundations vec-tionary is to infer the coordinates for a moral axis $\boldsymbol{m}$ for each of the five moral foundations.

Second, we defined the *observed relevance* (s_i) of an eMFD word as its association with a target moral foundation, already available in eMFD through a crowdsourcing procedure. Specifically, each word's observed relevance can be obtained by merging two pieces of key information from the eMFD: the probability and the sentiment scores of the word. In the eMFD wordlist, each word was assigned a probability score (ranging from 0 to 1) for its relevance to a specific moral foundation through crowd-sourced annotations (Hopp *et al.* 2021). Additionally, the eMFD captures the sentiment score of each MFD word per foundation, which ranges from −1(most negative sentiment, associated with moral vices) to +1 (most positive sentiment, associated with moral virtues). For each eMFD word, we merged the magnitude of the probability score and the sign of the sentiment valence to operationalize observed relevance, where s_i, p_i, and v_i represent the observed relevance, the probability magnitude, and the sentiment sign of a word i in the eMFD, correspondingly. As an example, consider the eMFD word "*kill*" with a Care/Harm foundation probability score of 0.40 and a sentiment score of −0.70. We incorporated the negative sign of the sentiment score ("−1") into the probability score, yielding an observed relevance of −0.40 for "*kill*". When constructing the moral foundations vec-tionary, we incorporated the "sign" (positive or negative) of each word's "sentiment score" available in eMFD but ignored its numeric value. This decision was based on the fact that, unlike eMFD's probability scores, the sentiment scores are derived from VADER, a simple rule-based lexicon (Hutto and Gilbert 2014). In other words, unlike probability scores, the eMFD sentiment scores have not yet undergone systematic crowdsourcing-based validation. Therefore, compared with the magnitude of "sentiment," the "sign," or the information on valence, is likely to be more robust and valid. The results reported in the performance comparison Section 4.2 demonstrate that our decision produced measurements better aligned with the benchmark "ground truth" than eMFD. That said, we acknowledge that future studies could benefit from considering different methods to derive "observed relevance" based on specific research needs, such as modifying the computation or using other seed dictionaries besides eMFD.

While the observed relevance s_i was directly obtained from human annotations during the development of the eMFD, it cannot be directly repurposed to uncover the moral axes in the vector space. To do so, we defined the *analytical relevance* of an eMFD word regarding a moral foundation, denoted as $\hat{s}_i$, as the scalar projection of that word's embedding on a particular moral axis. For example, for an eMFD word i in a 300-dimension vector space, its analytical relevance, $\hat{s}_i$, represents its scalar projection onto the moral axis $\boldsymbol{m}$, where word vector $\boldsymbol{w_i} = (w_{i,1}, w_{i,2}, \ldots, w_{i,300})$, $\boldsymbol{m} = (m_1, m_2, \ldots, m_{300})$. We can derive $\hat{s}_i$ as follows in Equation (1), where θ is the angle between vector $\boldsymbol{w_i}$ and $\boldsymbol{m}$, and $\|\cdot\|$ represents the 2-norm. To simplify the calculation, we normalized the word vector $\boldsymbol{w_i}$. Therefore, the analytical relevance $\hat{s}_i$ is essentially the cosine similarity between eMFD word vector $\boldsymbol{w_i}$ and the presumed moral axis $\boldsymbol{m}$. Cosine similarity is a standard measure in semantic vector space, which uses the cosine value of the angle between word vectors to measure their relevance (An *et al.* 2018; Mikolov *et al.* 2013). In this case, the analytical relevance $\hat{s}_i$ captures how closely the word is aligned to the moral axis $\boldsymbol{m}$.

$$\hat{s}_i = \cos\theta \cdot \|\boldsymbol{w_i}\| = \frac{\boldsymbol{w_i} \cdot \boldsymbol{m}}{\|\boldsymbol{w_i}\| \cdot \|\boldsymbol{m}\|} \cdot \|\boldsymbol{w_i}\| . \tag{1}$$

Next, we define the error e_i between the observed relevance s_i and the analytical relevance $\hat{s}_i$ for a specific word i, as indicated by Equation (2). This formulation helps define an objective function for the optimization algorithm, which seeks to identify the coordinates for the moral axis $\boldsymbol{m}$ that minimizes the summation of errors for all eMFD words, as defined in Equation (3), where N is the number of words considered. In this study, we used the nonlinear optimization solver Ipopt 0.6.5 for estimation. Interested researchers are welcome to experiment with other optimization algorithms for their domain applications.

$$e_i = (\hat{s}_i - s_i)^2 \tag{2}$$

$$\min \sum_{i=1}^{N} e_i. \tag{3}$$

Finally, we added Equation (4) as a constraint to normalize the moral axis $\boldsymbol{m}$.

$$\|\boldsymbol{m}\| = 1. \tag{4}$$

To summarize, the proposed model includes an objective function (3) with three equality constraints defined in (1), (2), and (4). The key output is the estimated coordinates of the moral axis $\boldsymbol{m}$. The main input data includes the eMFD wordlist, along with their vector representations and observed relevance values. The pipeline is implemented in Python 3.10 (for data processing) and Julia 1.6.2 (for optimization). Specifically, JuMP 0.21.10 and Ipopt 0.6.5 are used to solve the optimization problem, which was completed within 120 seconds for a 300-dimensional vector space and a total of 3,270 eMFD words.

Applying the vec-tionary framework requires several key decisions, including the selection of a validated dictionary, word embeddings, and optimization algorithms. Researchers have the flexibility to make these decisions to address their specific research needs. In our study, we used the eMFD as the seed dictionary due to its extensive validation through crowdsourcing. Regarding word embedding, we chose word2vec for its straightforward structure (i.e., two-layer neural networks), ease of use, and popularity. However, other validated dictionaries and next-generation word embeddings (e.g., OpenAI's text embedding models) can be considered as they become available. For the optimization algorithm, we selected the nonlinear optimization solver, Ipopt, to infer moral axes that minimize the sum of the L2 norm of errors between analytical and observed relevance, see Equation (2). Several reasons prompted us to choose Ipopt: (a) the high dimensionality of the word embedding space makes global optimization algorithms inefficient and overly complex, and (b) the need to handle the unit norm constraint, as shown in Equation (4), points to interior-point-based algorithms. We encourage future research to explore alternative optimization algorithms that better align with their specific analytical tasks. For example, the L1 norm can be applied when sparsity is a desired feature for the moral axis, or when outliers should carry less weight. Other solvers, such as BARON and NLopt, can be used when the model size is tractable. Finally, in our case study, for observed relevance, we combined eMFD's probability scores with the sign of sentiment scores while ignoring their magnitude. Although this procedure produced superior measurements than eMFD against our crowdsourced benchmark data, researchers might find alternative ways to calculate observed relevance, such as factoring in the numeric values of sentiment scores, that are more appropriate for their specific applications. Validation is the key to evaluating such decisions.

3.2. Three Measurement Metrics

Compared to the dictionary approaches (e.g., eMFD), the moral foundations vec-tionary has the advantage of providing multiple metrics to capture more nuanced aspects of moral content in textual data. Beyond measuring the magnitude of moral content in a text (*Strength*), the vec-tionary also captures the degree of expressed virtue versus vice for a particular moral foundation (*Valence*). Additionally, our approach also measures the degree of variance among the virtue-vice moral axis for a particular type of moral content (*Ambivalence*), to capture moral conflict such as the co-existence of both virtue- and vice-related expressions in a document. Neither the Valence nor the Ambivalence metric is available in previous moral foundations dictionaries. This expanded range of metrics not only enriches the scope of analysis for moral content but also bolsters the utility of vec-tionaries in the computational analysis of message features.

Specifically, the first metric, *Strength*, is denoted as the averaged *absolute values* of word-level projections (i.e., cosine similarities) of a document, as indicated by Equation (5), where n represents the number of words in a document, and θ_i is the angle between the vector representation of word i and the obtained moral axis $\boldsymbol{m}$. The *Strength* score ranges from 0 to 1, with larger values indicating a stronger moral foundation-specific relevance in the document regardless of valence. Here, we note that the *Strength* metric of the Moral Foundations Vec-tionary is conceptually similar to eMFD's probability, as both are designed to measure the magnitude of morally relevant content in texts. However, they utilize different methodological designs, and our empirical evidence in Section 4.2 demonstrates that *Strength* outperforms eMFD's probability.

The second metric, *Valence*, calculates the averaged word-level cosine similarities, ranging from -1 to 1, see Equation (6). It evaluates whether a document leans toward one side of a target moral foundation, with a positive *Valence* score indicating the use of virtue-dominated moral expressions and a negative *Valence* score indicating vice dominance. Virtue- versus vice-related moral expressions might nullify each other. For example, if a conservative tweet talks about "*saving immigrants' lives*" in the context of "*threatening local community safety*," it appeals to both the virtue and vice aspects of the Care/Harm moral foundation. This simultaneous use of care-related and harm-related cues can cancel each other out, resulting in a low moral valence for the text despite its strong but conflicting moral expression. We also note that although eMFD's sentiment scores are meant to capture a similar construct, its calculation is based on a separate sentiment lexicon that measures general sentiment positivity or negativity and determines the emotional tone of the message. Furthermore, eMFD's sentiment scores have not undergone systematic crowdsourcing, unlike its probability scores. In contrast, the *Valence* scores of the Moral Foundations Vec-tionary are based on geometric projections that utilize the same amount of moral "signals" from the crowdsourced eMFD probability scores as well as word embeddings.

As for *Ambivalence*, the last metric is a novel contribution of the vec-tionary approach and is designed to assess the co-presence of both moral virtue and vice-related expression in texts. It calculates the variance of word-level cosine similarities, ranging from 0 to 1, as defined in Equation (7). This metric captures the variability in word-level moral cues in a document. A higher *Ambivalence* score may indicate an expression of greater moral conflict. For instance, in the same tweet example given above, despite the overall valence being low, the resulting high *Ambivalence* score shows the tweet appealing to both sides of the *Care/Harm* moral foundation when it comes to its stance on immigrants (additional example tweets are available in Section A of the Supplementary Material). This new metric would allow researchers to examine how people express conflicting moral sentiments in short social media posts and other texts when they discuss controversial issues.

$$S = \frac{\sum_{i=1}^{n} |\cos\,\theta_i|}{n} \tag{5}$$

$$V = \frac{\sum_{i=1}^{n} \cos\,\theta_i}{n} \tag{6}$$

$$A = \frac{\sum_{i=1}^{n} (\cos\,\theta_i - V)^2}{n}. \tag{7}$$

The moral foundations vec-tionary offers several advantages over previous measurement strategies. First, it establishes moral axes based on a large, validated set of 3,270 dictionary words rather than relying

https://doi.org/10.1017/pan.2025.6 Published online by Cambridge University Press

on a limited number of seed words (see two examples in Table B1 in Section B of the Supplementary Material). This enhances comprehensiveness and robustness. Second, the moral foundations vec-tionary recognizes that different words may contribute differently to a moral dimension, unlike previous methods that assume equal contributions of seed words. Third, constructing moral axes with varying word weights presents mathematical challenges, as uncovering coordinates in a high-dimensional vector space is a non-trivial problem. We took advantage of a non-linear optimization algorithm to extract the maximal amount of moral signals from all eMFD words, along with their corresponding weights. Lastly, the moral foundations vec-tionary extends beyond the 3,270 eMFD words to harness additional moral signals from other words in a given corpus through word embeddings. In our case, the vec-tionary captures signals from over 300 million words and phrases, significantly broadening the spectrum of moral content that can be analyzed within any text. To the best of our knowledge, this represents the first attempt in the literature on computational analyses of moral content to enhance an established dictionary with word embeddings and a formal optimization algorithm. Next, we present empirical evidence comparing the performance of the moral foundations vec-tionary with eMFD benchmarked on a "ground truth" tweets dataset in the context of a politicized public health crisis in which diverse moral discussions have been widely merged.

3.3. An Efficient and Easy-to-Use Python Package

To facilitate other researchers in their analyses, we are releasing a Python package called `vMFD` that implements our method. Our package includes the moral charges of over 300 million words and phrases calculated using the proposed approach. All three metrics above have been implemented. The code is open-sourced on GitHub.[1]

The package is easy to install and works out of the box. It has been indexed in PyPI, the official third-party package repository for Python. Installing `vMFD` needs a single command: `pip install vMFD`, and analyzing text messages only requires a few lines of code. The package is also highly efficient. Our tests show that processing one million tweets with `vMFD` on a modern laptop (e.g., M1 MacBook Pro) takes about seven minutes using a single processor.

4. Model Validation and Performance Comparison

We validated the performance of moral foundations vec-tionary against the eMFD on a benchmark dataset of COVID-19 tweets annotated for the moral *Strength* through a crowdsourcing procedure.

4.1. Annotators, Training, and Annotation Procedure

4.1.1. Annotation Platform

We developed a crowdsourcing system that implements the pairwise comparison task built based on the open-sourced "All Our Ideas" project (Salganik and Levy 2015), also known as the "wiki-surveys" (www.allourideas.org), see details of our customized platform in Section C of the Supplementary Material. For each moral foundation, we created two tasks, one measuring the virtue aspect of the foundation and the other the vice aspect (e.g., one question on *care* and the other on *harm* for the *Care/Harm* foundation), gathering human annotators' moral judgments on the same set of tweets, consistent with prior practices (Hoover *et al.* 2020).

The statistical rationale for the pairwise comparison task and the procedures to estimate per-message moral scores from the annotation results are detailed elsewhere (Salganik and Levy 2015). In a nutshell, the system constructs an opinion matrix based on respondents' selected tweets from each pair (see an example task interface in Figure C1 in Section C of the Supplementary Material) and estimates the latent score for each tweet through Bayesian inference and a hierarchical probit model. Conceptually,

[1] https://github.com/ZeningDuan/vMFD

the resulting latent score for a particular tweet, ranging between 0 and 100, can be interpreted as its likelihood of outperforming a randomly chosen tweet for a randomly selected annotator: a minimum of 0 indicates consistent loss, while a maximum of 100 means the tweet would always win. For instance, when assessing the virtue of *care*, a tweet that reads, "*After ousting a dictator, members of Sudan's resistance committees are now helping to fight the Covid-19 pandemic*," receiving a score of 96, suggests that for a random annotator, this tweet would be estimated to outperform a randomly selected tweet 96% of the time. Finally, foundation by foundation, we were able to construct an overall ranking of all the annotated tweets based on the estimated moral *Strength* scores from the crowdsourcing system.

4.1.2. Annotators Recruitment and Training

For each moral foundation, to produce sufficient data density (Carlson and Montgomery 2017; Hopp *et al.* 2021), we ensured that at least 70% of the tweets in the stimuli pool will be evaluated by at least 15 annotators (for calculation details, refer to Section D of the Supplementary Material).

Annotators were recruited from the Prolific platform, and each of them was assigned two tasks: one focused on the virtue dimension and the other on vice, each task involving at least 25 pairs of tweets. To avoid potential order effects, we randomized the sequence of these two annotation tasks. Furthermore, we matched this sample to census distributions on five key demographic variables: gender, age, ideological affiliation, education, and race (descriptive statistics details see Section E of the Supplementary Material).

Each annotator focuses on one randomly assigned moral foundation. Before tasks, they are invited to an online training module, and only those who pass are eligible to proceed to annotation tasks (training materials in Section F of the Supplementary Material). After further screening to exclude annotators who failed the test or were timed out, we retained a total of 3,473 qualified annotators in the analytical sample, informed consent was obtained.

4.1.3. Stimuli Corpus for Annotation

We collected tweets from June 15, 2020 to July 12, 2020, through Twitter's COVID-19 firehose API. Since Twitter's original search query includes non-English terms, we applied core 25 keywords (see Table B2 in Section B of the Supplementary Material) to further filter the corpus to make our dataset more focused. After preprocessing, this procedure resulted in a total of 2,285,379 unique English tweets. Tweets contain moral content (Hoover *et al.* 2020), while the overall prevalence could be low, thus we stratified sampled tweet stimuli by eMFD scores. To ensure sufficient variance in our stimuli corpus, for each moral foundation, we randomly selected 800 tweets from the strata with the highest eMFD scores for virtue and vice. Then, we added 400 tweets with low scores across all five foundations as control. This sampling strategy yielded 2,000 unique tweets per moral foundation (see Sections G and H of the Supplementary Material for details).

4.2. Performance Comparison Based on the Rank-Biased Overlap (RBO) Method

We assessed the moral foundations vec-tionary by comparing its outputs with the eMFD scores, using crowdsourced human annotations as the "ground truth." Given that the three approaches (vec-tionary, eMFD, and human annotation) used different scales for moral relevance scores, our focus was on comparing the rankings of the 2,000 tweets per foundation by these methods.

For the moral foundations vec-tionary, the ranking was built based on its *Strength* scores. Conceptually, they reflect the relevance of moral foundation in texts and are, therefore, comparable to eMFD's probability scores. Regarding the crowdsourced benchmark dataset, we built the ranking by summing up the square of both the virtue and vice scores for each tweet. To simplify the notations, hereafter we refer to the moral foundations vec-tionary as vec-tionary and the transformed crowdsourced scores aggregating virtue and vice as C.S. for brevity.

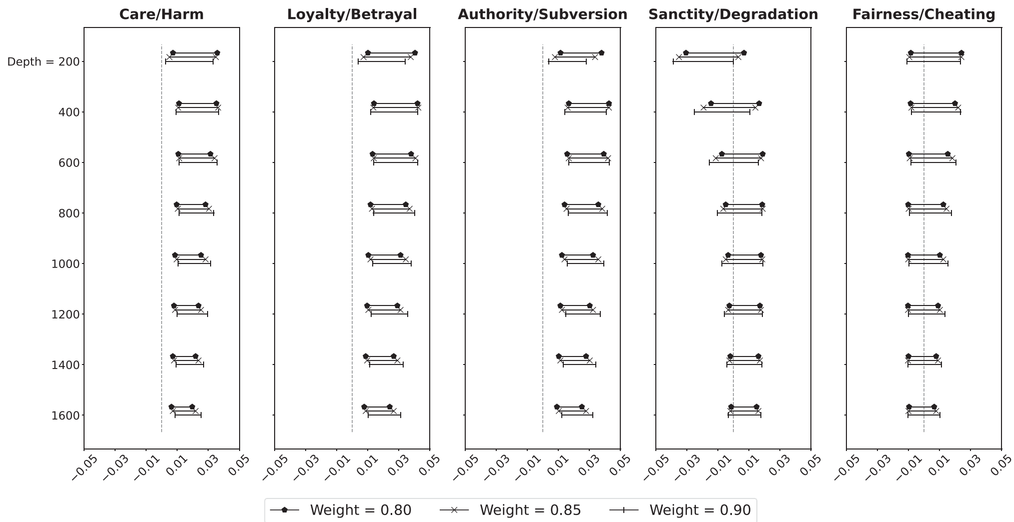

Figure 2. Difference in RBO similarity scores by moral foundation.

We define the similarity between ranking i and ranking j as $R_{i,j}$. We used the RBO method measure $R_{vec-tionary,CS}$ and $R_{eMFD,CS}$ accordingly and then computed their difference $R_{vec-tionary,CS} - R_{eMFD,CS}$ (see the RBO Equation in Section I of the Supplementary Material). RBO, first introduced by Webber and colleagues, is a continuous measure that quantifies the similarity between two ranked lists (Webber, Moffat, and Zobel 2010). It has been used in many fields (Ng and Taneja 2019; Urman, Makhortykh, and Ulloa 2022) and has been shown to be more sensitive to the positions of overlapping items compared to other similarity measures. RBO takes into account both the depth of overlap (i.e., how many items are shared between the two lists) and the rank positions of the overlapping items (i.e., how close the overlapping items are to the top of the lists). It assigns more weights to items that are ranked higher than lower, aligning with our interest in measuring rank changes and assigning greater weight to the top of the list of stimuli tweets compared to those occurring further down. RBO provides adjustable parameters to systematically explore how similarities might change as the researcher places more weight to items at the top of the two rankings. In our case, RBO allows for a robustness check to assess the measurement performance of the moral foundations vec-tionary versus the eMFD while varying the degree to which tweets with stronger moral cues should dominate the calculation of similarities. We calculated $R_{vec-tionary,CS}$ and $R_{eMFD,CS}$ with varying weights and depths in the ranking comparison (see details in Table I1 in Section I of the Supplementary Material). This approach follows similar practices in existing studies (Ng and Taneja 2019; Urman *et al.* 2022). To quantify estimation uncertainty for the difference between $R_{vec-tionary,CS}$ and $R_{eMFD,CS}$, we employed bootstrapping (resamples = 5,000, with replacement) to estimate the 95% confidence intervals (CIs).

Results show that irrespective of varying depths and weights, the moral foundations vec-tionary rankings consistently showed higher similarities with the crowdsourcing benchmark rankings than the eMFD for three moral foundations: *Care/Harm*, *Loyalty/Betrayal*, and *Authority/Subversion* (see Figure 2). Regarding the *Sanctity/Degradation* foundation, as the *depth* parameter increased, the moral foundations vec-tionary showed a tendency to outperform the eMFD, albeit falling short of reaching the conventional threshold for statistical significance. Furthermore, these two methods did not significantly differ with regards to the *Fairness/Cheating* foundation. To facilitate interpretation, in Section A of the Supplementary Material, we provide exemplar tweets where the moral foundations vec-tionary produced more accurate results than the eMFD.

To better quantify measurement improvement regarding the three moral foundations where the moral foundations vec-tionary outperformed the eMFD, we calculated the metric Percentage

https://doi.org/10.1017/pan.2025.6 Published online by Cambridge University Press

Table 1. Performance comparison for the Care/Harm moral foundation while varying weight and depth values.

		RBO similarities			PPI of vec-tionary
		Vec-tionary vs. C.S.	eMFD vs. C.S.	Vec-tionary vs. eMFD	over eMFD (%)
	Weight = .80	.16	.12	.26	33.24
Depth 200	Weight = .85	.13	.09	.21	39.50
	Weight = .90	.11	.07	.17	47.50
	Weight = .80	.28	.24	.40	17.60
Depth 400	Weight = .85	.23	.19	.35	22.00
	Weight = .90	.19	.15	.30	27.60
	Weight = .80	.38	.34	.50	11.05
Depth 600	Weight = .85	.31	.27	.44	14.82
	Weight = .90	.27	.23	.39	18.50
	Weight = .80	.44	.41	.56	8.26
Depth 800	Weight = .85	.40	.36	.52	10.11
	Weight = .90	.33	.29	.45	13.88
	Weight = .80	.50	.47	.60	6.47
Depth 1000	Weight = .85	.44	.41	.56	8.26
	Weight = .90	.38	.34	.50	11.05
	Weight = .80	.53	.51	.63	5.61
Depth 1200	Weight = .85	.50	.47	.60	6.47
	Weight = .90	.42	.38	.53	9.18
	Weight = .80	.57	.55	.66	4.77
Depth 1400	Weight = .85	.53	.51	.63	5.61
	Weight = .90	.47	.44	.58	7.36
	Weight = .80	.61	.59	.69	3.96
Depth 1600	Weight = .85	.57	.54	.66	4.77
	Weight = .90	.50	.47	.60	6.47

Note: Vec-tionary = moral foundations vec-tionary; C.S. = crowdsourced benchmark scores; PPI = percentage performance increase.

Performance Increase (PPI), see Equation (8).

$$\text{PPI} = \frac{R_{vec-tionary,\,CS} - R_{eMFD,\,CS}}{R_{eMFD,\,CS}}. \tag{8}$$

We calculated the pairwise similarity scores between vec-tionary, eMFD, and crowdsourced benchmark, foundation by foundation, along with the PPI scores contrasting vec-tionary's performance with that of eMFD. Table 1, which consists of four columns, provides an illustrative example. The first three columns represent comparisons among the vec-tionary, the C.S., and the eMFD, while the last column indicates the percentage increase of vec-tionary over eMFD, respectively. For instance, the first row shows that for the *Care/Harm* foundation, when *weight* was set as .80, and *depth* as 200, the RBO similarities are as follows: $R_{vec-tionary,\,CS} = .16$, $R_{eMFD,\,CS} = .12$, and $R_{vec-tionary,\,eMFD} = .26$. Furthermore, a PPI score of 33.24% indicates that the Moral Foundations Vec-tionary improves the measurement of *Care/Harm* appeals by 33.24%, compared to the eMFD, when benchmarked against crowdsourced

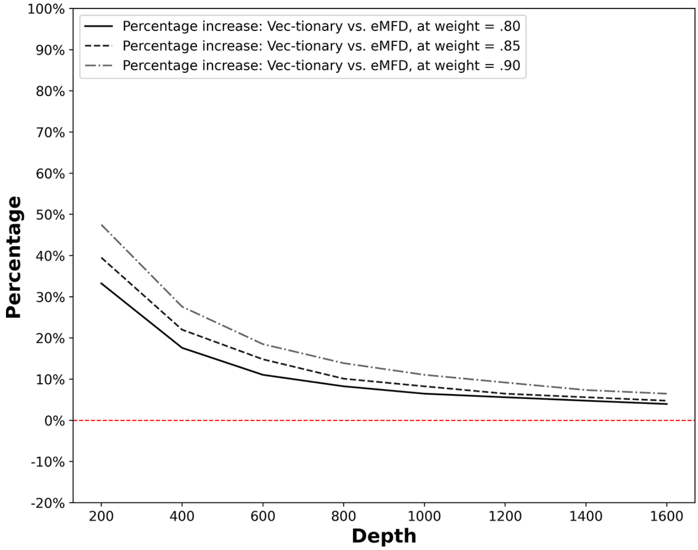

Figure 3. Performance gain of the moral foundations vec-tionary: Care/Harm.

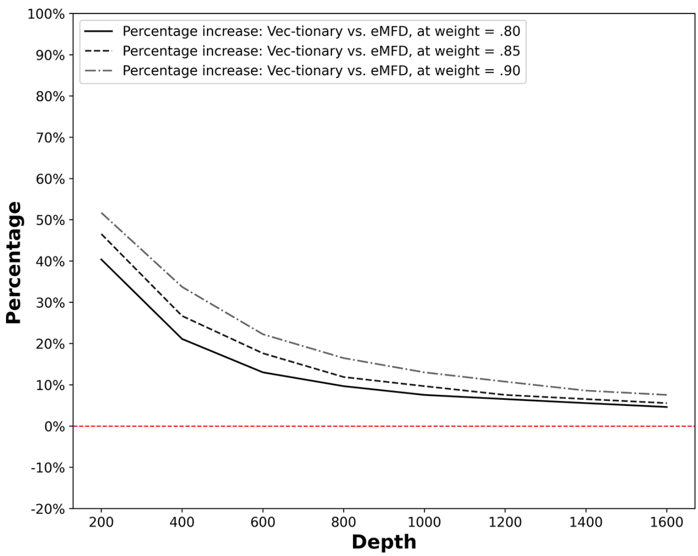

Figure 4. Performance gain of the moral foundations vec-tionary: Loyalty/Betrayal.

human annotations. For the remaining foundations, we have summarized their calculations in Tables J2 to J5, which can be found in Section J of the Supplementary Material.

In Figures 3,4,5, we visualized the PPI scores for three moral foundations (*Care/Harm*, *Authority/Subversion*, and *Loyalty/Betrayal*). The results show that the moral foundations vec-tionary tends to outperform the eMFD more with lower *depth* values and higher *weight* values. This pattern suggests that the moral foundations vec-tionary is particularly sensitivein capturing stronger moral cues within texts because the combination of lower *depth* and higher *weight* would correspond to prioritizing top-ranked tweets for a given moral foundation, in similarity calculation (see Section J of the Supplementary Material for more details). This property of our moral foundations vec-tionary is arguably desirable, as many research applications would focus on social media posts that contain strong and clear

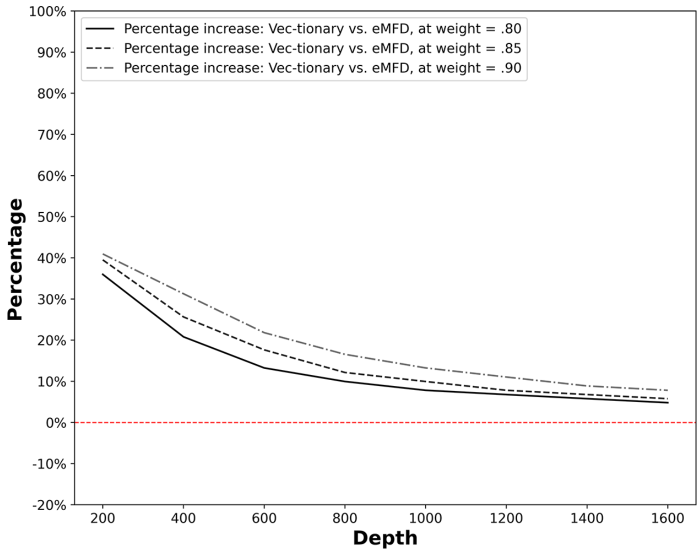

Figure 5. Performance gain of the moral foundations vec-tionary: Authority/Subversion.

moral signals. We also constructed SEMAXIS for comparison and found that vec-tionary consistently outperforms it, as detailed in Section J of the Supplementary Material.

5. An Application of the Moral Foundations Vec-tionary

Public opinion scholars have long been intrigued by the theoretical question of which specific message features, such as moral content, function as "triggers" for increased online retransmission (Brady *et al.* 2017; Brady *et al.* 2019). In this study, we aim to illustrate how the moral foundations vec-tionary can effectively identify moral content within tweets. Additionally, we explored its ability to predict the number of retweets, surpassing the eMFD scores, after controlling for common covariates. Furthermore, we sought to assess whether additional measurement metrics, namely, moral *Valence* and *Ambivalence*, which are not directly available in the eMFD, can account for unique variances beyond moral *Strength*, and offer deeper conceptual insights into the nuances of moral content. To facilitate interpretation, we provide exemplar tweets with high or low scores on each of the three different metrics, *Strength*, *Valence*, and *Ambivalence*, in Section A of the Supplementary Material.

Prior to fitting the models, we applied standard text-preprocessing procedures to the corpus of COVID-19 tweets (details available in Section H of the Supplementary Material). Given our interest in predicting the number of retweets as a case study to illustrate the usefulness of measures from the moral foundations vec-tionary, we guaranteed that each tweet in our corpus had an equal chance to accrue retweets by applying an identical 14-day moving window. Additionally, we incorporated metadata such as account verification status and expressed emotion valence as control variables. The main outcome, the number of retweets, is a count variable with skewed distribution characterized by over-dispersion and a high proportion of zeros (78.69% of the total dataset). Therefore, we employed the Zero-Inflated Negative Binomial (ZINB) regression to examine the relationships between moral content and retweeting.

Figure 6 summarizes model specification for the six models that we fit to assess the predictive power of moral content: Model 1 was the baseline model with only metadata and expressed emotions; Model 2 added eMFD scores to Model 1, and Model 3 in turn added moral *Strength* scores from the moral foundations vec-tionary to Model 2; Model 4 and 5 respectively added moral *Valence* and *Ambivalence* scores to Model 3, and Model 6 was the "kitchen sink" full model incorporating all predictors previously

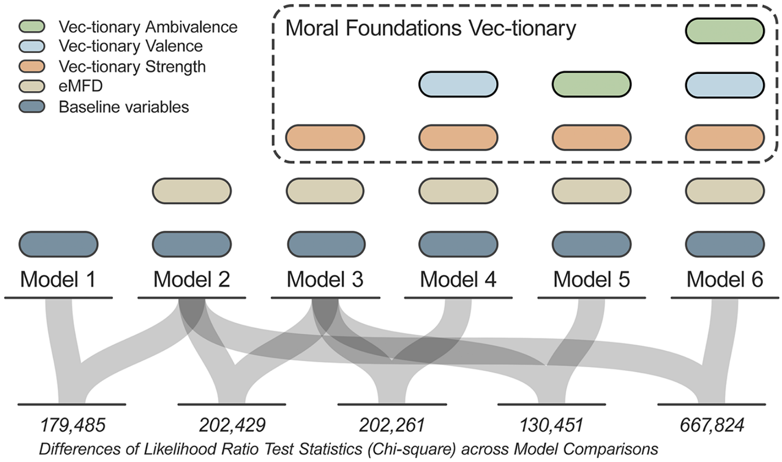

Figure 6. Model specification and comparison.

mentioned. The addictive structure of these models allows us to unpack whether metrics from the moral foundations vec-tionary can account for unique variances in predicting the number of retweets through a series of likelihood ratio tests (LRTs) (Lewis, Butler, and Gilbert 2011). We also assessed changes in Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as complementary evidence. To address potential multicollinearity concerns, we assessed the Variance Inflation Factor (VIF) for each model and observed VIF values within the acceptable range of 1.01–9.40, indicating no significant multicollinearity issues given the context of our large dataset and complex models.

We found that each metric (i.e., foundation-specific Strength, Valence, and Ambivalence scores) from the moral foundations vec-tionary incrementally accounted for unique variance in the zero-inflated negative binomial regression model predicting retweet counts, above and beyond the moral probability scores from the eMFD and metadata (see Figure 6). Echoing prior research (Brady *et al.* 2017), Model 2 incorporating eMFD probability scores significantly improved model fit over the baseline Model 1 ($\chi^2(10) = 179{,}485; \Delta_{AIC} = -179{,}465, \Delta_{BIC} = -179{,}338$), demonstrating that moral content measured through eMFD significantly predicted retweeting. Model 3, adding moral *Strength* measures from the moral foundations vec-tionary, explained additional variances than Model 2 ($\chi^2(10) = 202{,}429; \Delta_{AIC} = -202{,}409, \Delta_{BIC} = -202{,}283$), suggesting that the moral foundations vec-tionary captured unique moral signals beyond eMFD. Next, we assessed whether the two additional metrics from the moral foundations vec-tionary, i.e., moral *Valence* and *Ambivalence*, enhanced predictive power beyond moral *Strength* and the eMFD measures. Model 4 and 5, adding *Valence* and *Ambivalence* scores, respectively, both outperformed Model 3 ($\chi^2(10) = 202{,}261$ for Model 4; $\Delta_{AIC} = -202{,}241, \Delta_{BIC} = -202{,}115; \chi^2(10) = 130{,}451$, for Model 5; $\Delta_{AIC} = -130{,}431, \Delta_{BIC} = -130{,}305$). Lastly, comparing the full model (Model 6) with Model 2, we found a significant model fit improvement by incorporating all three metrics from the moral foundations vec-tionary ($\chi^2(30) = 667{,}824; \Delta_{AIC} = -667{,}764, \Delta_{BIC} = -667{,}385$).

The regression results, as shown in Table 2 (full version available in Section K of the Supplementary Material), are threefold. First, we found that each foundation-specific Strength score differentially predicted the outcome. Strength scores capture the magnitude of relevance to a particular moral foundation irrespective of valence. Specifically, in the full model (i.e., Model 6), tweets expressing Care/Harm ($\beta = 0.050$, 95% CI: 0.048–0.052) and Loyalty/Betrayal ($\beta = 0.059$, 95% CI: 0.058–0.060) are associated with more retweets; in contrast, tweets with higher Fairness/Cheating ($\beta = -0.187$, 95% CI: −0.188 to −0.185) and Sanctity/Degradation ($\beta = -0.268$, 95% CI: −0.269 to −0.266) predicted fewer retweets. Unlike previous research that lumps all five moral foundations together in predicting retweeting (Brady *et al.* 2017), our results underscore the importance of taking a moral pluralism

Table 2. Model outputs (only showing Poisson regression) in evaluating vec-tionary's predictive capabilities.

	Dependent variable: count number of retweets			
	Model 3	Model 4	Model 5	Model 6
Care/Harm (Strength)	0.145	0.005	0.234	0.050
	(0.144, 0.146)	(0.004, 0.006)	(0.233, 0.235)	(0.048, 0.052)
Fairness/Cheating (Strength)	0.002	−0.022	−0.083	−0.187
	(0.001, 0.003)	(−0.023, −0.021)	(−0.084, −0.081)	(−0.188, −0.185)
Loyalty/Betrayal (Strength)	−0.009	0.050	−0.049	0.059
	(−0.010, −0.008)	(0.049, 0.050)	(−0.050, −0.048)	(0.058, 0.060)
Authority/Subversion (Strength)	0.013	−0.001	0.009	−0.005
	(0.012, 0.013)	(−0.002, −0.0001)	(0.008, 0.011)	(−0.006, −0.004)
Sanctity/Degradation (Strength)	−0.018	−0.076	−0.096	−0.268
	(−0.019, −0.017)	(−0.077, −0.075)	(−0.097, −0.095)	(−0.269, −0.266)
Care/Harm (Valence)		−0.120		−0.106
		(−0.121, −0.118)		(−0.107, −0.104)
Fairness/Cheating (Valence)		−0.053		−0.114
		(−0.054, −0.052)		(−0.115, −0.113)
Loyalty/Betrayal (Valence)		0.003		0.011
		(0.002, 0.004)		(0.010, 0.013)
Authority/Subversion (Valence)		−0.045		−0.053
		(−0.046, −0.044)		(−0.054, −0.052)
Sanctity/Degradation (Valence)		−0.056		−0.146
		(−0.057, −0.055)		(−0.147, −0.145)
Care/Harm (Ambivalence)			−0.109	−0.033
			(−0.110, −0.108)	(−0.034, −0.031)
Fairness/Cheating (Ambivalence)			0.118	0.185
			(0.116, 0.119)	(0.184, 0.186)
Loyalty/Betrayal (Ambivalence)			0.047	−0.034
			(0.046, 0.048)	(−0.035, −0.033)
Authority/Subversion (Ambivalence)			−0.014	−0.022
			(−0.015, −0.013)	(−0.023, −0.021)
Sanctity/Degradation (Ambivalence)			0.089	0.171
			(0.088, 0.090)	(0.170, 0.173)
Metadata (incl.)				
Emotion (incl.)				
eMFD probability (incl.)				
Constant	3.489	3.453	3.467	3.436
	(3.488, 3.490)	(3.452, 3.454)	(3.466, 3.468)	(3.435, 3.437)

(continued)

Table 2. Continued.

	Dependent variable: count number of retweets			
	Model 3	Model 4	Model 5	Model 6
Observations	2,285,379			
Log Likelihood	−42,398,612	−42,297,481	−42,333,387	−42,165,915
AIC	84,797,304	84,595,063	84,666,873	84,331,949
BIC	84,797,810	84,595,695	84,667,505	84,332,708

Note: The italic values indicate the 95% confidence intervals for the corresponding Poisson regression coefficients reported in regular font.

perspective (Graham *et al.* 2018) and demonstrate the theoretical value of unpacking moral appeals' foundation-specific effects on online message diffusion.

The metric of Valence evaluates the directional leaning of a text along the moral foundation axis. Notably, we uncovered evidence suggesting a "virtue penalty," where tweets scoring higher in expressing moral virtues showed a disadvantage in accruing retweets. This "penalty" of virtue expression holds true in four out of the five moral foundations tested (i.e., Care: $\beta = -0.106$, 95% CI: −0.107 to −0.104; Fairness: $\beta = 0.114$, 95% CI: −0.115 to −0.113; Authority: $\beta = -0.053$, 95% CI: −0.054 to −0.052; Sanctity: $\beta = -0.146$, 95% CI: −0.147 to −0.145). This interesting pattern suggests that Twitter users are more likely to retweet messages that highlight violations of moral principles, perhaps reflecting the social regulation function of morality as well as an evolutionary sensitivity toward moral transgressions (Haidt 2012). To the best of our knowledge, these results represent the first large-scale demonstration of such "virtue penalty" in online message retransmission.

As for Ambivalence scores, this metric is a novel contribution of the vec-tionary approach and is designed to assess the co-presence of both moral virtue and vice-related expression in texts. A high ambivalence score can be interpreted to suggest the expression of moral conflict. For instance, in the context of COVID-19, such conflicted moral expressions are not uncommon (see Section A of the Supplementary Material for examples). This new metric would allow researchers to examine how people express conflicting moral sentiments in short social media posts when they discuss controversial issues such as COVID-19. In the added empirical results (see Table 2), we found that tweets that expressed a higher level of moral ambivalence regarding Fairness/Cheating ($\beta = 0.185$, 95% CI: 0.184 to 0.186) and Sanctity/Degradation ($\beta = 0.171$, 95% CI: 0.170 to 0.173) actually garner more retweets, after controlling for Strength and Valence scores. In contrast, higher Ambivalence scores related to the other three foundations showed negative, albeit much weaker, associations (i.e., Care/Harm: $\beta = -0.033$, 95% CI: −0.034 to −0.031; Loyalty/Betrayal: $\beta = -0.034$, 95% CI: −0.035 to −0.033; Authority/Subversion: $\beta = -0.022$, 95% CI: −0.023 to −0.021). Although it is beyond the scope of the current study to pin down the exact mechanisms that could explain these foundation-specific effects related to moral ambivalence, our speculation is that Tweets highlighting moral dilemmas and uncertainties related to fairness (e.g., equal access to vaccines vs. prioritizing vulnerable populations) and sanctity (e.g., "disgusting" ingredients in vaccines vs. protecting the body from viruses) during the pandemic might be notably attention-grabbing and shareworthy. At least, these findings demonstrate the value of the Ambivalence metric in capturing a unique aspect of moral expression when discussing controversial issues, sometimes dubbed as "wicked problems" where uncertainties and polarizing reactions in both factual understanding and value judgments are prevalent (Head 2022; Lilleker and Stoeckle 2021).

In summary, our findings consistently demonstrate improved model performance when incorporating the three metrics from the moral foundations vec-tionary. Though conceptually similar to the eMFD moral scores, the moral *Strength* metric from the moral foundations vec-tionary accounted for unique variances in predicting retweeting beyond the eMFD. Furthermore, the two additional metrics, moral *Valence* and *Ambivalence*, offered unique conceptual values and explanatory power. Therefore, researchers can benefit from adopting the three distinct metrics that the moral foundations vec-tionary provides for a multifaceted assessment of in-text moral content.

https://doi.org/10.1017/pan.2025.6 Published online by Cambridge University Press

6. Discussion

We introduce a novel computational approach to develop *vec-tionaries*, a measurement tool outperforming conventional dictionaries in extracting and measuring message features from texts. In this paper, we focus on moral content as a case study, due to growing scholarly interest in studying the roles of moral content in public opinion, political engagement and persuasion, online communicative behaviors, among others (Chen *et al.* 2024; Feinberg and Willer 2013; Graham *et al.* 2013; Solovev and Pröllochs 2023; Zhou *et al.* 2022). The moral foundations vec-tionary draws from extensive methodological literature on measuring moral content, notably the eMFD based on crowdsourcing (Hopp *et al.* 2021) and the DDR method based on word embeddings (An *et al.* 2018; Garten *et al.* 2018). In constructing the moral foundations vec-tionary, we employed nonlinear optimization algorithms to estimate moral axes in a semantic vector space by merging crowdsourced moral ratings from the eMFD with established word embeddings.

The moral foundations vec-tionary stands out in several ways. First, its architectural framework allows outputting an array of metrics, including *Strength*, *Valence*, and *Ambivalence*, to quantify distinctive aspects of moral content—a noteworthy expansion broadening the scope of available measures from existing various moral foundations dictionaries (i.e., eMFD), as detailed in the Section 3.2. Moral *Strength* captures the presence and magnitude of a particular type of moral content in a text, collapsing the virtue and vice dimensions of the corresponding moral foundation. Our validation analyses through the RBO analyses, refer to the Section 4.2, have largely confirmed the superiority of the moral *Strength* measure from the moral foundations vec-tionary, benchmarked against crowdsourced human annotations. Furthermore, the *Valence* measure assesses the predominant moral sentiment by taking the net difference between expressed virtue and vice for a given moral foundation, whereas *Ambivalence* measures the variance along the foundation-specific virtue-vice axis—for example, higher values of *Ambivalence* could be interpreted as indicating higher moral conflict, i.e., mentioning both virtue and vice. In the reported application in Section 5 predicting tweet retransmission, we not only reaffirmed our previous validation results by showing the unique variances accounted for by moral *Strength* scores but also underscored the significance of incorporating *Valence* and *Ambivalence*—these results remain valid even after controlling for eMFD scores, expressed emotions, and other baseline predictors. Taken together, the moral foundations vec-tionary not only yields better measurements for moral *Strength*, but also opens new avenues for researchers to explore, particularly regarding moral ambiguity and conflict through the *Ambivalence* metric since discussions about virtue and vice often co-occur within the same message.

Another notable advantage of vec-tionaries is that, unlike traditional dictionary-based methods that consider only a limited set of keywords, vec-tionaries encompass all available words within a given text. This distinction is essential because conventional dictionaries often risk invoking false negative errors—incorrectly indicating the absence of a moral foundation—when context-specific moral signals are contained in words absent from the dictionary's word list. In contrast, vec-tionaries employ nonlinear optimization to harness continuous ratings from the full list of eMFD words while incorporating additional moral signals from other words of a given text beyond the eMFD list. In the context of studying moral content, this property becomes especially valuable when researchers are interested in analyzing moral content within short-form texts such as social media posts (Brady *et al.* 2017; Zhou *et al.* 2022), where signals are scarce. Directly applying the eMFD to short-form social media posts might not yield accurate measurements because the eMFD was originally developed for measuring long-form texts such as news stories. The original authors of the eMFD have also emphasized this limitation (Hopp *et al.* 2021). The moral foundations vec-tionary can help scholars interested in studying naturally occurring moral expressions online to identify and capture a much wider range of message instances with high external validity (see an analysis conducted to identify moral words captured by the moral foundations vec-tionary but missed by the eMFD word list in Section L of the Supplementary Material).

The last notable advantage of vec-tionaries is contextual adaptability captured through word embeddings. Conventional dictionary-based methods often neglect context-specific nuances. In contrast, vec-tionaries allow the selection of word embeddings tailored to specific contexts. For instance, researchers can substitute the default general-purpose word embeddings (e.g., word2vec, GloVe) with embeddings tailored to the specific context or application. The model can also incorporate word embeddings from fine-tuned large language models such as Generative Pre-trained Transformers (GPTs) as they become available.

This study emphasizes the importance of validation by benchmarking crowdsourced data. We developed a protocol to crowdsource human annotations of moral content within short-form texts, taking insights from the pairwise comparison paradigm (Carlson and Montgomery 2017; Salganik and Levy 2015). Given the documented difficulty in measuring moral content following conventional manual coding procedures (Weber *et al.* 2021), our crowdsourcing protocol fills a critical gap in the literature and can be used to construct "ground truth" datasets for moral content in other applications. Our validation results confirmed better performance of the moral foundations vec-tionary for three (i.e., Care/Harm, Authority/Subversion, and Loyalty/Betrayal) out of the five moral foundations tested, particularly for tweets containing stronger moral signals. For the remaining two moral foundations, the measurement accuracy of the moral foundations vec-tionary was on par with the eMFD. We encourage future research to replicate this documented between-foundation variation in performance in other contexts and to further investigate underlying mechanisms. Taken together, these results suggest that the moral foundations vec-tionary is a valid tool for measuring moral content from texts. That said, we do not suggest that the moral foundations vec-tionary should replace the eMFD, rather, researchers are welcome to use and test this new tool as a complementary resource to existing methods. Additionally, while vec-tionaries are cost-effective and quick when using pre-validated dictionaries like eMFD and LIWC, they do not necessarily outperform fine-tuned BERT-style models, which require substantial human and computational resources. Vec-tionaries offer the advantage of transparency and easier interpretability, making them particularly valuable in academic and applied settings where unpacking the "blackbox" of computational methods is emphasized.

Measuring and classifying text features, such as moral content, often serve as the initial step for statistical analyses that help social scientists explain other outcomes. A common oversight involves ignoring measurement errors, which can lead to biased estimators and invalid confidence intervals in downstream regression analyses. Labels from computational models such as LLMs, BERT, or vec-tionaries, as used in our study, can be imperfect and deviate from the true labels. Recent methodological advances point to promising ways to address such measurement errors from computational labels for message features, including a design-based supervised learning estimator (Egami *et al.* 2024). We encourage future research to consider this approach or other methods to mitigate potential biases and incorporate measurement errors in outputs from vec-tionaries.

Through demonstrating the validity and utility of the moral foundations vec-tionary, our aspiration is to illustrate the conceptual basis and methodological framework of a novel method that combines validated dictionaries with word embeddings to measure latent message features such as moral appeals, which we call vec-tionaries. Since researchers can follow the procedures outlined in this paper, we encourage interested researchers to develop their own vec-tionaries to measure other latent message features (e.g., emotional appeals, incivility, linguistic sophistication, politicizing frames) across languages and contexts. Three key steps to construct vec-tionaries are worth bearing in mind: first, find a validated dictionary with wordlists and weights measuring the targeted latent feature; second, select a set of word embeddings, either general-purpose or context-specific; and finally, specify an appropriate optimization algorithm to construct semantic axes aligned with the desired latent feature(s). By following these steps, the constructed vec-tionary can yield continuous measurements for the targeted message features, including strength, valence, and ambivalence. In closing, we reiterate the importance of adopting an agnostic approach and conducting validation tests before using the constructed vec-tionary for substantive analyses (Grimmer *et al.* 2022).

Acknowledgements. We are grateful for the valuable insights and feedback from Christopher Lucas, Jiaxin Pei, Qijia Ye, and colleagues at SICSS-Penn and various conferences, including IC2S2, ICA, Media & Morality meeting, and APSA, which significantly contributed to our work. We also thank Micol Federica Tresoldi, Chenghui Li, Ye Wang, Luyu Xu, Doug Hemken, Jiyoun Suk, Yini Zhang, Michael W. Wagner, and Dhavan V. Shah for their support and comments at different stages. Finally, We deeply appreciate our editor and anonymous reviewers for their valuable time and tremendous efforts.

Competing Interests. The authors declare none.

Funding Statement. Our project receives support from the John S. and James L. Knight Foundation (Award Number: MSN231314). We also received funding from the Wisconsin Alumni Research Foundation (WARF) to S.Y. (MSN231886) and K.C. (AAH2162) and a WARF Accelerator Big Data Challenge Grant from the same funder awarded to S.Y. and Z.D. (MSN275569).

Data Availability Statement. Replication datasets and code are available on Harvard Dataverse: https://doi.org/10.7910/DVN/YITNSV. Please note that to protect privacy and to comply with X's terms of service regarding data sharing, only tweet IDs are made available for analyses reported in Section 5 of this manuscript. Interested readers need to "re-hydrate" these IDs to access the full text of analyzed tweets. Open-source Python software to create and implement "vec-tionaries" introduced in this article is available on GitHub: https://github.com/ZeningDuan/vMFD.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2025.6.

References

An, J., H. Kwak, and Y.-Y. Ahn. 2018. "SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2450–2461. Melbourne: Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1228.

Araque, O., L. Gatti, and K. Kalimeri. 2022. "LibertyMFD: A Lexicon to Assess the Moral Foundation of Liberty." In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, 154–160. New York: Association for Computing Machinery. https://doi.org/10.1145/3524458.3547264.

Brady, W. J., J. A. Wills, D. Burkart, J. T. Jost, and J. J. Van Bavel. 2019. "An Ideological Asymmetry in the Diffusion of Moralized Content on Social Media Among Political Leaders." *Journal of Experimental Psychology: General* 148 (10): 1802.

Brady, W. J., J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel. 2017. "Emotion Shapes the Diffusion of Moralized Content in Social Networks." *Proceedings of the National Academy of Sciences* 114 (28): 7313–7318.

Carlson, D., and J. M. Montgomery. 2017. "A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts." *American Political Science Review* 111 (4): 835–843.

Chen, K., Z. Duan, and S. J. Kim. 2024. "Uncovering Gender Stereotypes in Controversial Science Discourse: Evidence from Computational Text and Visual Analyses Across Digital Platforms." *Journal of Computer-Mediated Communication* 29 (1): zmad052.

Clifford, S. 2014. "Linking Issue Stances and Trait Inferences: A Theory of Moral Exemplification." *The Journal of Politics* 76 (3): 698–710.

Clifford, S., and J. Jerit. 2013. "How Words do the Work of Politics: Moral Foundations Theory and the Debate Over Stem Cell Research." *The Journal of Politics* 75 (3): 659–671.

Egami, N., M. Hinck, B. Stewart, and H. Wei. 2024. "Using Imperfect Surrogates for Downstream Inference: Design-Based Supervised Learning for Social Science Applications of Large Language Models." *Advances in Neural Information Processing Systems* 36: 68589–68601.

Feinberg, M., and R. Willer. 2013. "The Moral Roots of Environmental Attitudes." *Psychological Science* 24 (1): 56–62.

Frimer, J., J. Haidt, J. Graham, M. Dehghani, and R. Boghrati. 2017. "Moral Foundations Dictionaries for Linguistic Analyses, 2.0." Unpublished Manuscript. www.jeremyfrimer.com/uploads/2/1/2/7/21278832/summary.pdf.

Garten, J., et al. 2018. "Dictionaries and Distributions: Combining Expert Knowledge and Large Scale Textual Data Content Analysis: Distributed Dictionary Representation." *Behavior Research Methods* 50: 344–361.

Graham, J., et al. 2013. "Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism." *Advances in Experimental Social Psychology*, 47: 55–130. San Diego, CA: Elsevier.

Graham, J., J. Haidt, M. Motyl, P. Meindl, C. Iskiwitch, and M. Mooijman. 2018. "Moral Foundations Theory: On the Advantages of Moral Pluralism Over Moral Monism." *Atlas of Moral Psychology* 211: 222.

Graham, J., J. Haidt, and B. A. Nosek. 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations." *Journal of Personality and Social Psychology* 96 (5): 1029.

Grimmer, J., M. E. Roberts, and B. M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton, NJ: Princeton University Press.

Haidt, J. 2012. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. New York: Vintage.

Head, B. W. 2022. "The Rise of 'Wicked Problems'—Uncertainty, Complexity and Divergence." In *Wicked Problems in Public Policy: Understanding and Responding to Complex Challenges*, 21–36. Cham, Switzerland: Springer.

Hoover, J., et al. 2020. "Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment." *Social Psychological and Personality Science* 11 (8): 1057–1071.

Hopkins, D. J., and G. King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–247.

Hopp, F. R., J. T. Fisher, D. Cornell, R. Huskey, and R. Weber. 2021. "The Extended Moral Foundations Dictionary (eMFD): Development and Applications of a Crowd-Sourced Approach to Extracting Moral Intuitions from Text." *Behavior Research Methods* 53: 232–246.

Hutto, C., and E. Gilbert. 2014. "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media*, 8, 216–225. https://doi.org/10.1609/icwsm.v8i1.14550.

Jung, J.-H. 2020. "The Mobilizing Effect of Parties' Moral Rhetoric." *American Journal of Political Science* 64 (2): 341–355.

Kaplan, J. T., A. Vaccaro, M. Henning, and L. Christov-Moore. 2023. "Moral Reframing of Messages About Mask-Wearing During the COVID-19 Pandemic." *Scientific Reports* 13 (1): 10140.

Kozlowski, A. C., M. Taddy, and J. A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class Through Word Embeddings." *American Sociological Review* 84 (5): 905–949.

Kwak, H., J. An, E. Jing, and Y.-Y. Ahn. 2021. "FrameAxis: Characterizing Microframe Bias and Intensity with Word Embedding." *PeerJ Computer Science* 7: e644.

Lewis, F., A. Butler, and L. Gilbert. 2011. "A Unified Approach to Model Selection Using the Likelihood Ratio Test." *Methods in Ecology and Evolution* 2 (2): 155–162.

Liang, H., Y. M. M. Ng, and N. L. Tsang. 2023. "Word Embedding Enrichment for Dictionary Construction: An Example of Incivility in Cantonese." *Computational Communication Research* 5 (1): 1.

Lilleker, D. G., and T. Stoeckle. 2021. "The Challenges of Providing Certainty in the Face of Wicked Problems: Analysing the UK Government's Handling of the COVID-19 Pandemic." *Journal of Public Affairs* 21 (4): e2733.

Mikolov, T., W.-T. Yih, and G. Zweig. 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*, 746–751. Atlanta: Association for Computational Linguistics. https://aclanthology.org/N13-1090.

Mucciaroni, G. 2011. "Are Debates About "Morality Policy" Really About Morality? Framing Opposition to Gay and Lesbian Rights." *Policy Studies Journal* 39 (2): 187–216.

Ng, Y. M. M., and H. Taneja. 2019. "Mapping User-Centric Internet Geographies: How Similar are Countries in their Web Use Patterns?" *Journal of Communication* 69 (5): 467–489.

Pennington, J., R. Socher, and C. D. Manning. 2014. "Glove: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Rodman, E. 2020. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis* 28 (1): 87–111.

Sagi, E., and M. Dehghani. 2014. "Measuring Moral Rhetoric in Text." *Social Science Computer Review* 32 (2): 132–144.

Salganik, M. J., and K. E. Levy. 2015. "Wiki Surveys: Open and Quantifiable Social Data Collection." *PloS One* 10 (5): e0123483.

Solovev, K., and N. Pröllochs. 2023. "Moralized Language Predicts Hate Speech on Social Media." *PNAS Nexus* 2 (1): pgac281.

Theocharis, Y., P. Barberá, Z. Fazekas, S. A. Popa, and O. Parnet. 2016. "A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates." *Journal of Communication* 66 (6): 1007–1031.

Urman, A., M. Makhortykh, and R. Ulloa. 2022. "The Matter of Chance: Auditing Web Search Results Related to the 2020 US Presidential Primary Elections Across Six Search Engines." *Social Science Computer Review* 40 (5): 1323–1339.

Webber, W., A. Moffat, and J. Zobel. 2010. "A Similarity Measure for Indefinite Rankings." *ACM Transactions on Information Systems (TOIS)* 28 (4): 1–38.

Weber, R., et al. 2021. "Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions." In W. van Atteveldt & T.-Q. Peng (Eds.), *Computational Methods for Communication Science*, 39–59. New York, NY: Routledge.

Yang, F., and S. Yang. 2023. "Effects of Moral Frames Within Vaping Prevention Messages on Current Smokers' Support for Electronic Cigarette Regulations." *Journal of Health Communication* 28 (7): 412–424.

Zhou, A., et al. 2022. "Moral Foundations, Ideological Divide, and Public Engagement with US Government Agencies' COVID-19 Vaccine Communication on Social Media." *Mass Communication and Society* 27 (4): 739–764.

Cite this article: Duan, Z., Shao, A., Hu, Y., Lee, H., Liao, X., Suh, Y. J., Kim, J., Yang, K.-C., Chen, K. and Yang, S., (2025). Constructing Vec-tionaries to Extract Message Features from Texts: A Case Study of Moral Content. *Political Analysis*, 1–21. https://doi.org/10.1017/pan.2025.6