UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Richie, R., et al. (2019). Predicting High-Level Human Judgment Across Diverse Behavioral Domains. *Collabra: Psychology,* 5(1): 50. DOI: https://doi.org/10.1525/collabra.282

ORIGINAL RESEARCH REPORT

Predicting High-Level Human Judgment Across Diverse Behavioral Domains

Russell Richie, Wanling Zou and Sudeep Bhatia

Recent advances in machine learning, combined with the increased availability of large natural language datasets, have made it possible to uncover semantic representations that characterize what people know about and associate with a wide range of objects and concepts. In this paper, we examine the power of word embeddings, a popular approach for uncovering semantic representations, for studying high-level human judgment. Word embeddings are typically applied to linguistic and semantic tasks, however we show that word embeddings can be used to predict complex theoretically- and practically- relevant human perceptions and evaluations in domains as diverse as social cognition, health behavior, risk perception, organizational behavior, and marketing. By learning mappings from word embeddings directly onto judgment ratings, we outperform a similarity-based baseline and perform favorably compared to common metrics of human inter-rater reliability. Word embeddings are also able to identify the concepts that are most associated with observed perceptions and evaluations, and can thus shed light on the psychological substrates of judgment. Overall, we provide new methods and insights for predicting and understanding high-level human judgment, with important applications across the social and behavioral sciences.

Keywords: judgment; semantic memory; machine learning; word embeddings

Introduction

Recent advances in machine learning, combined with the increased availability of large natural language datasets, have made it possible to uncover semantic representations that characterize what people know about and associate with a wide range of objects and concepts. In this paper, we examine the power of word embeddings, a popular approach for uncovering semantic representations, for studying high-level human judgment. Word embeddings are typically applied to linguistic and semantic tasks, however we show that word embeddings can be used to predict complex theoretically- and practically- relevant human perceptions and evaluations in domains as diverse as social cognition, health behavior, risk perception, organizational behavior, and marketing. By learning mappings from word embeddings directly onto judgment ratings, we outperform a similarity-based baseline and perform favorably compared to common metrics of human inter-rater reliability. Word embeddings are also able to identify the concepts that are most associated with observed perceptions and evaluations, and can thus shed light on the psychological substrates of judgment. Overall, we provide new methods and insights for predicting and understanding high-level human judgment, with important applications across the social and behavioral sciences.

People are constantly perceiving, judging and evaluating entities in the world, on the qualities that these entities possess. They may consider, for example, whether a food item is nutritious, whether a political candidate is competent, whether a consumer brand is exciting, or whether the work of an occupation is significant. Such judgments influence every sphere of life, determining the social, professional, consumer, and health outcomes of individuals, as well as the political and economic makeup of our societies. It is thus of critical importance to social and behavioral scientists to develop predictive and explanatory models of human judgment. To have good empirical coverage and practical utility, such models must apply to naturalistic objects and concepts, i.e., the vast range of entities people encounter every day and have rich knowledge about. They should be able to quantify what people know about these entities, and specify how people map this knowledge onto the diverse array of complex judgments they make on a day-to-day basis.

We show how it is possible to quantify knowledge and predict complex judgment with a high degree of accuracy, naturalism, and generality. Our approach relies on *word embeddings*, a popular class of models in machine learning that use the statistics of word distribution in language to derive high-dimensional vectors for words and phrases (see Lenci, 2018 for review). These vectors represent

Department of Psychology, University of Pennsylvania, US
Corresponding author: Russell Richie (drrichie@sas.upenn.edu)

semantic knowledge, so that similar or related words have vectors that are closer to each other in the underlying semantic space. Word embedding-based semantic representations are a useful tool for many practical natural language processing and artificial intelligence applications (Turney & Pantel, 2010). However, they also mimic aspects of human semantic cognition, and thus can be used to study how people learn, represent, and manipulate the meanings of words (Mandera, Keuleers, & Brysbaert, 2017). Indeed, word embeddings are able to accurately predict human responses in semantic judgment tasks, such as tasks involving assessments of word similarity, relatedness, and association (Bruni, Tran, & Baroni, 2014; Hill, Reichart, & Korhonen, 2015; Hofmann, Biemann, Westbury, Murusidze, Conrad, & Jacobs, 2018; Levy, Bullinaria, & McCormick, 2017). They are also useful for modeling other phenomena related to memory, and recent work has used this approach to study priming and lexical access as well as free association, semantic search, and list recall (Bhatia, 2017; Healey & Kahana, 2016; Hills, Jones, & Todd, 2012; Jones, Kintsch, & Mewhort, 2006; Mandera et al., 2017). For this reason, word embeddings are becoming increasingly popular in fields like psycholinguistics, cognitive psychology, and cognitive neuroscience. More recently, scientists have begun to extend word embeddings beyond linguistic and semantic judgment. Many areas of psychology involve associative processing, and measurements of word vector similarity can be used to specify the associations that determine peoples' responses. Based on this insight, researchers have found that word embeddings also predict certain association-based probability judgments, social judgments, and consumer judgments (Bhatia, 2017, 2018; Caliskan, Bryson, & Narayanan, 2017; Garg, Schiebinger, Jurafsky, & Zou, 2018).

We find that the structure of knowledge captured by word embedding-based semantic representations can also be applied to study a very wide range of complex human judgments, including judgments that are not easily captured by association-based measures of vector similarity. More specifically, we find that with some training data, it is possible to learn a mapping from word-embeddings space to the judgment domain in consideration, and subsequently make accurate predictions for nearly any entity in that domain. This learnt mapping can also be used to identify the concepts that are most related to the judgment, and thus understand the most important psychological factors for the judgment. To illustrate the broad applicability of this method, we use it to study fourteen types of judgment across seven different domains in the behavioral sciences. These judgments involve naturalistic entities, such as food items, consumer goods, personality traits, job occupations, brands, and public figures, and we examine judgments for these entities on key theoretical dimensions, as identified by prior research.

From representation to judgment

There are a number of influential methods used by behavioral scientists to quantify representations for the purpose of studying perceptions, evaluations, and other types of judgments for naturalistic entities. These methods can all be considered to be *psychometric* in that they require human measurements, typically in the form of numeric ratings on scales (McRae, Cree, Seidenberg, & McNorgan, 2005; Shepard, 1980; Slovic, 1987). Although psychometric techniques have been applied quite successfully across a variety of domains, they typically yield representations of target entity knowledge that are impoverished compared to what people know about the targets. As these approaches rely on explicit participant measurement, is also often costly to collect the information needed to study complex real-world judgment. Of course, it is also impossible to use psychometric techniques to quantify how uncovered representations for judgment targets relate to the hundreds of thousands of other objects and concepts that are involved in the mental lives of individuals and are not necessarily judgment targets.

Thus, a technique is needed which cheaply delivers rich, high-dimensional knowledge representations for a large number of objects and concepts, which can then be used to model judgments. Fortunately, such a technique can be found in *word embeddings*. As with some psychometric approaches, such as multidimensional scaling (Shepard, 1980), word embedding techniques rely on object similarity to uncover entity representations. However, it is not explicit participant similarity ratings or categorization judgments that are used in this analysis, but rather similarity in language, so that pairs of words that occur in similar linguistic contexts are given similar representations. With the recent availability of large-scale natural language online data, as well as new computational resources for analyzing this data, it is possible to use contextual similarity in language to uncover vector representations for all of the words (and corresponding objects and concepts) used in everyday discourse. The dimensions of the vectors are analogous to the latent dimensions uncovered using psychometric methods, but the vectors themselves are typically far richer than what could be obtained through survey-based techniques. Most word embedding models have representations for hundreds of thousands of objects and concepts, with each representation involving hundreds of underlying dimensions.

Our proposed approach follows from established psychometric techniques, in which vector representations for entities are used as independent variables in a regression predicting the judgment in consideration. Thus, we also use our word embeddings-based vector representations as inputs into a regression. Due to the relatively high-dimensionality of the word embeddings we use various regularized regression algorithms, and evaluate our model's performance in terms of its ability to predict out-of-sample participant judgments. Intuitively, this approach uses participant judgments for some entities to learn a *mapping* from the high-dimensional semantic space to the judgment dimension in consideration. When a new entity is presented to be judged, it applies its learned mapping to the semantic vector for the new entity to predict participant judgments for that entity. Some prior work has used this approach to generate word norms in psycholinguistic studies (Hollis,

Westbury, & Lefsrud, 2017), and to study risk perception (Bhatia, 2019) and brand judgments (Bhatia & Olivola, 2018). In principle, this approach can be applied to any type of human judgment, as long as judgment targets are in the word embeddings vocabulary. It will be successful as long as participant judgment also rely on elements of semantic knowledge represented within the word embeddings space.

We also compare our approach against a baseline model, which involves calculations of *vector similarity* on the word embeddings space. As similar vectors are semantically related or associated, it is possible to predict the judgment of an entity on a given dimension by calculating how close its vector is to words describing high vs. low ends of the judgment dimension. This measure of relative distance can then be passed through a linear transformation to predict judgments on the same scale as responses elicited from participants. Nearly all applications of word embeddings to predict human semantic and linguistic judgment, memory phenomena such as priming and free association, as well as high-level social judgment and probability judgment, involve calculations of relative vector similarity. Although, this approach performs very well, it is unlikely that it will be successful if the judgments in consideration involves more complex transformations than just simple associations with a set of predetermined judgment-relevant words. That said, our proposed mapping approach can be seen as an extension of the baseline vector similarity approach that uses training data to infer the regions of the semantic space that are most associated with the judgment dimension. As we illustrate below, this property of the mapping approach makes it suitable for inferring, in a bottom-up manner, the psychological substrates of the judgment in consideration.

Figure 1 shows a hypothetical vector semantic space, and illustrates how both our primary (mapping) approach and our baseline (similarity) approach can be applied to representations in this space to predict judgment. Additional details of these two approaches are provided in the methods and materials section.

Results

Summary of Data

Our goal in this paper is to evaluate the broad applicability of word embeddings for predicting high-level judgment. Thus, we have chosen seven diverse behavioral domains, spanning a range of disciplines in the behavioral sciences, including social cognition, health behavior, risk perception, organizational behavior, and marketing. For each of the seven domains we have identified two theoretically relevant judgment dimensions, that have been the focus of considerable prior research, and are considered to play a foundational role in human behavior. We have also identified 200 different naturalistic entities for each of these seven domains, selected to cover the range of variation within a domain (see section Item Generation and Table S1 in the Supplementary Information (SI) for details). As we wished to compare the accuracy of our method across these domains, we obtained participant ratings of these entities on the two corresponding judgment dimensions in a single large study. Overall, our study involved 140,000 participant judgments, along 14 judgment dimensions, for nearly 1,400 distinct entities (see SI for all items), spanning seven semantic domains. The judgment dimensions, domains, items, participant instructions, and various implementation details for this study and for the resulting analysis, have been pre-registered at https://osf.io/t7qyb/. A summary of judgment dimensions and items we consider, along with

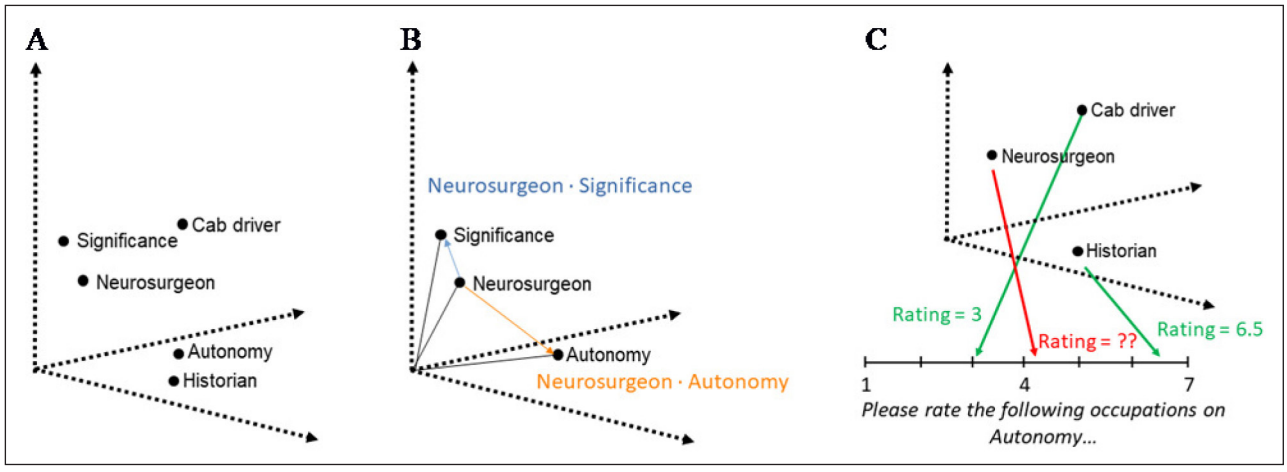

Figure 1: (A) An illustration of a hypothetical three dimensional semantic space with representations for three occupations (neurosurgeon, historian, and cab driver) and two properties (autonomy and significance) on which the occupations are to be judged. **(B)** An illustration of the vector similarity approach. We use the dot product between the vector for an occupation and the vector for a property to predict the association between the occupation and propery. In this example, neurosurgeon is predicted to be more associated with significance than with autonomy. **(C)** An illustration of the vector mapping approach. Here we use participant data to learn the mapping from the vector space to the rating dimension in consideration. We train a supervised model to predict autonomy ratings for cab driver and historian based on their vectors, and use the resulting model to predict the autonomy rating for neurosurgeon given its vector.

Table 1: The judgments of the current study, along with relevant fields, example applications, sample items, and classic references in which these judgments are measured and studied.

Judgments	Relevant Fields	Example Applications	Sample Items	Classic References
Masculinity and femininity of traits	Social psychology; personality psychology	Gender roles	arrogant, gentle, sociable	Bem (1974)
Dread-inducement and unknowability of potential risk sources	Behavioral economics; risk analysis; public policy	Risk behaviors	marijuana, tsunami, hackers	Slovic (1987)
Warmth and competence of people	Social psychology; behavioral economics	Interpersonal behavior from dating to voting	Bill Clinton, Adolf Hitler, Mother Teresa	Rosenberg et al. (1968); Fiske et al. (2002) Cuddy et al. (2002)
Taste and nutrition of foods	Health psychology; public health policy	Dietary behavior; public health	carrots, tiramisu, celeriac	Raghunathan, Naylor, and Hoyer (2006)
Significance and autonomy of occupations	Industrial-organizational psychology; labor economics	Career choices; job satisfaction	cab driver, neurosurgeon, historian	Hackman and Oldham (1976)
Sincerity and excitement of brands	Marketing; consumer psychology; industrial-organizational psychology	Purchasing behavior; organization-public relations	Home Depot, Comedy Central, ING Direct	Aaker (1997)
Hedonic and utilitarian value of goods	Marketing; consumer psychology; psychology of motivation	Purchasing and consumption behavior	chips, vest, hammer	Batra and Ahtola (1990)

example fields and applications and relevant references, is provided in **Table 1**.

In line with our preregistered analysis plan, our main analysis uses a pre-trained word embedding model – word2vec – obtained using the skip-gram technique (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), applied to a very large dataset of Google News articles. This space has 300-dimensional vectors for 3 million words and short phrases, and we selected our judgment stimuli to ensure that they were present within the word2vec vocabulary. However we also replicate our basic results with a diverse array of other embedding models trained on various corpora.

Predictive Accuracy

We first evaluated the predictive accuracy of our method for the participant ratings in our data. This initial analysis pertains to predictive accuracy for average participant judgments (i.e. averages of the ratings made on each the fourteen judgment dimensions). As set out in our pre-registration, we tested the ability of a variety of (regularized) regression techniques, across a range of hyperparameters, for mapping pretrained 300-dimensional word embeddings to judgments. We evaluated predictive accuracy in a cross-validation exercise. Our methods and materials section, and Supplementary Information sections Cross-validation and Model Selection and Secondary Models, contain details of this procedure, as well as out-of-sample R^2 and root mean squared error (RMSE) scores for all model and hyperparameter combinations that we tested (see Figure S2). A range of models performed well, but we focus here on our best-performing model, a ridge regression with regularization hyperparameter λ set to 10. **Figure 2** shows, for each judgment dimension, scatterplots of actual judgments and predicted judgments, along with Pearson correlation coefficients, for this method. Each predicted judgment in the scatterplot was obtained by leave-one-out cross-validation (LOOCV): we trained our ridge regression model on the vectors for all but one judgment target, and then used the trained model to predict the rating for the left-out judgment target based on the target's vector. As can be seen in **Figure 2**, our approach was able to predict participant judgments with a high degree of accuracy, with an average correlation rate of .77 across the fourteen judgment dimensions, and all fourteen judgments yielding statistically significant positive correlations (all $p < 10^{-20}$).

We compared the vector mapping approach with the simpler, baseline approach that relies only on the relative similarity of a judgment target to words denoting high vs. low ends of a particular judgment dimension (see Methods and Materials and Supplemental Information Table S2 for more detail). This method also involves fits using leave-one-out cross validation, though training the baseline approach only involves learning a linear transformation from the measure of relative vector similarity to the response scale in consideration. We found that the average correlation using this method was .30, which is much lower than that obtained using the vector mapping method. Additionally, the similarity method yields significant correlations only for eleven out of the fourteen tests. Of course, such differences are to be expected, as the mapping method uses the underlying vector space in a much more flexible manner.

We then compared the predictive accuracy of both methods with human inter-rater reliability, as human inter-rater reliability is often thought to place an upper bound on machine performance (Grand, Blank, Pereira,

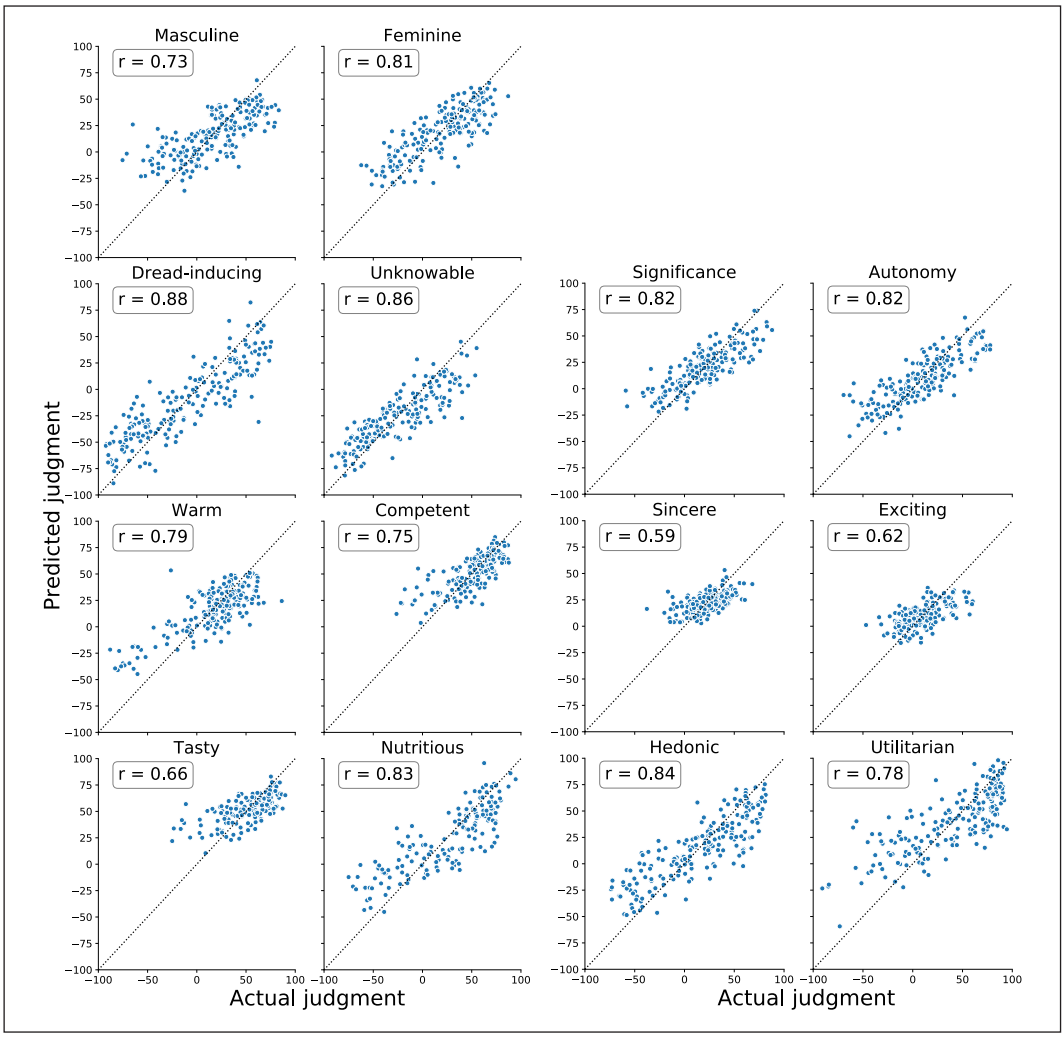

Figure 2: Scatterplots of actual judgments and predicted judgments using leave-one-out cross-validation for each judgment dimension.

& Fedorenko, 2018; Hill et al., 2015). For comparison to models predicting mean judgment ratings, we computed reliability two ways. First, we computed the inter-subject correlation (IS-r, Grand et al., 2018), which is the average correlation between one participant's ratings and the average of the rest. This is a commonly used metric in assessing word embeddings' ability to model semantic judgments (e.g., Grand et al., 2018) and is sometimes taken to place an upper bound on machine performance (Pilehvar & Camacho-Collados, 2018). This correlation came out to 0.60, whereas our main model surpassed this with an average correlation of 0.77 across judgments, meaning that our mapping method predicts mean human judgments better than do individual human judgments. However, given that our main model is predicting an average judgment rating with word embeddings that more or less constitute the 'average' of human knowledge reflected in word use, it may be more sensible to compare our models' performance to split-half reliability, or the correlation between the average of half the participants with the average of the other half of the participants. Thus, for each judgment dimension, we split participants into two sets, averaged judgment ratings within each set, computed the correlation between the averages, and repeated this process 100 times. The resulting split-half reliability in our judgments averaged across all judgment dimensions is .88, ranging from .69 for taste judgments to .97 for dread-inducing judgments. **Figure 3A** has similarity and mapping method predicted vs actual correlations under leave-one-out cross-validation, inter-subject correlations, and split-half reliabilities for every judgment dimension.

In addition to illustrating the accuracy of our proposed method relative to the baseline method and inter-rater reliability, **Figures 3A** also allows for a comparison of accuracy across judgment dimensions. Such a clean analysis of the differences across domains is possible since we randomly assigned subjects to judgment domains. As can be seen, while performance is good, there is a fair amount of variability in performance across dimensions: while risk source unknowability was predicted with a correlation of .86, food item taste was predicted with a correlation of .66. This variability is likely due to a variety of factors. For example, different judgments may vary in the extent to which they rely on knowledge reflected in language and thus word embeddings. Thus, for example, judgment-relevant attributes for taste may not be represented in our word embedding space. If this were the case, we would

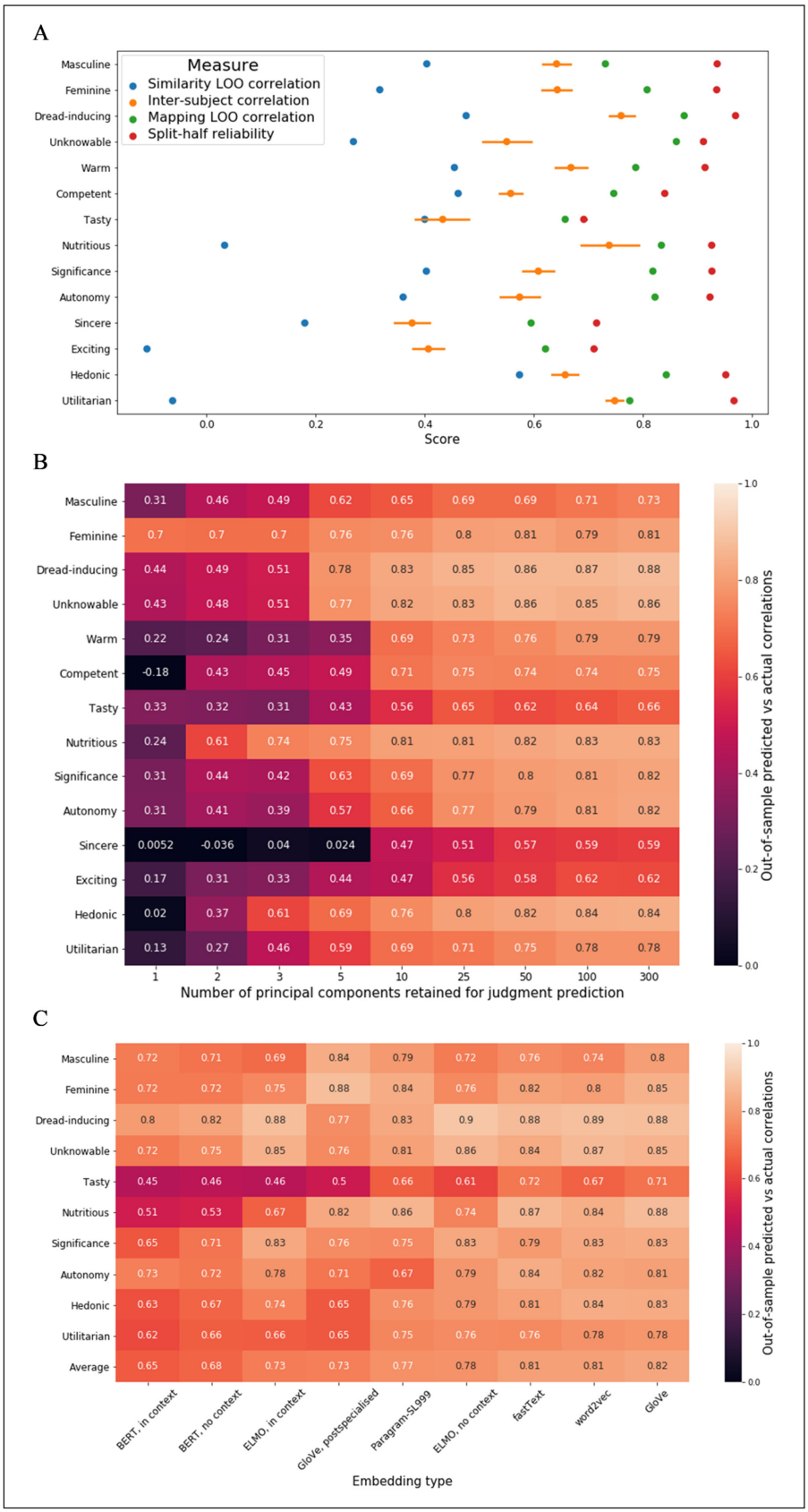

Figure 3: (A) Performance of the similarity baseline and mapping approach compared to inter-subject correlation (IS-r) and split-half reliability (error bars too small to visualize). The similarity baseline consistently performs under inter-subject correlation, while mapping outperforms this measure of reliability and approaches split-half reliability. **(B)** Pearson correlations between predicted and actual judgments for every judgment dimension and varying numbers of retained principal components. Judgment domains (brands, goods, traits, etc.) can be compressed to 5 to 25 principal components while preserving judgment prediction accuracy. **(C)** Pearson correlations between out-of-sample predicted and actual judgments for every judgment dimension except warmth and competence, for different word embedding models.

expect the accuracy of the baseline similarity approach (which also relies on the word embeddings space) to correlate with the accuracy of the proposed mapping approach. However, the relationship between the two approaches is very mild (r = .37, p = .19). One might also speculate that our judgment dimensions varied in their subjectivity: perhaps taste judgments are merely more subjective than other dimensions. Indeed, as **Figure 3A** shows, mapping model performance is strongly correlated with split-half reliability (r = .88, $p < 10^{-4}$), such that judgment domains where human raters disagreed more – taste and brand perceptions – were also more difficult for our model. In SI section "Comparing information use across judgment domains" and SI Figure S3, we further compare judgment dimensions, particularly how different judgment dimensions make use of the embeddings in similar and different ways. Interesting cross-domain connections arise, e.g., the similarity in embedding use between autonomy and hedonic value judgments, possibly due to a link between pleasure/happiness and self-determination.

Note that our approach can also be applied to individual-level judgments, thereby accommodating participant heterogeneity. As shown in the SI section Individual-level Modeling, we obtain average correlations of .52 for predicted vs. observed judgments, for the individual participants in each of our fourteen tests. These accuracy rates are lower than those obtained on the aggregate level, likely due to the fact that averaging participant ratings reduces variability in data. However, they are higher than accuracy rates obtained by applying the baseline approach to individual-level judgments, which generates average correlations of .21. As the baseline approach uses the same distances on the semantic space, for all participants, it cannot substantively accommodate participant heterogeneity (though this approach does allow for different participants to map vector similarities onto responses in different ways). To assess the individual-level models relative to inter-rater reliability, we again computed reliability two ways. First, we computed the average pairwise correlation between raters Hill et al. (2015). This correlation came out to 0.34, whereas our individual-level model predictions correlated with actual judgments at an average correlation of 0.53. We can also compare individual-level model accuracy with IS-r rates, since IS-r reflects the ability to predict an individual judgment from the mean of other judgments. As stated above, mean IS-r was .60, somewhat above our average individual-level model accuracy of .53. Figure S4 in the SI contains a pointplot like **Figure 3A** with these two measures of inter-rater reliability along with individual-level mapping and similarity models. Overall, for both average- and individual-level judgments, our model performs favorably in comparison to human inter-rater reliability, either exceeding inter-rater reliability or approaching it, depending on choice of inter-rater reliability metric.

Amount of Information Required for Prediction

A natural question for the present work is how much information in the 300-dimensional embeddings is actually required to represent our judgment targets, and hence predict our participants' judgments. To this end, we measured predictive accuracy through leave-one-out cross-validation with our primary ridge model ($\lambda = 10$) after reducing the embedding spaces with principal components analysis. Specifically, for each domain, we fit a principal component analysis on the training data design matrix (approximately 199 items, by 300 word2vec dimensions), applied the learned transformation to both the training and held-out data, discarded all but a certain number of initial principal components, and then tested how our ridge model trained on these dimension-reduced matrices predicted the held-out judgment. **Figure 3B** has predicted vs. actual Pearson correlations for every judgment dimension and number of retained principal components we tested. As can be seen, the 300-dimensional word embeddings can be compressed drastically – to <10% of their initial dimensionality – while preserving predictive performance, with only, on average, a 3-point drop in correlation strength when retaining only the first 25 PC's, and a 7-point drop when retaining only the first 10 PC's. This suggests that, within a domain, the representational space needed to predict the present kinds of judgments is much sparser than the space provided by word2vec. Theoretically, this shows that people may only be evaluating a relative handful of (latent) dimensions when making the kinds of judgments studied here. At the same time, that much of the information relevant to making these judgments is present in the initial principal components further validates previous claims that these 14 dimensions are core dimensions along which we represent objects in these seven domains (Aaker, 1997; Batra & Ahtola, 1990; Bem, 1974; Cuddy, Fiske, Glick, & Xu, 2002; Hackman & Oldham, 1976; Raghunathan, Naylor, & Hoyer, 2006; Rosenberg, Nelson, & Vivekananthan, 1968; Slovic, 1987). Practically, these results indicate that future applications of the tested method need not utilize all 300 dimensions, and that successful predictions can be obtained using standard, non-regularized regression methods in the behavioral sciences applied to 10- or 25-dimensional target spaces. What kinds of information the individual principal components represent is an important question for future research, but we believe these dimension-reduced spaces are a step towards more interpretable yet highly predictive models of judgment, as a modeler now has far fewer dimensions (10 to 25, vs 300) to examine or relate to interpretable psychological quantities.

We can also investigate the amount of information and the model complexity required for prediction by conducting a learning curve analysis, whereby we examine in- and out-of-sample prediction accuracy at increasing volumes of training data. SI section Learning Curve Analysis and Figure S7 has more details of this analysis, but in brief, we found that (1) prediction accuracy is still improving when training on almost all judgment targets, and (2) in-sample accuracy is consistently higher than out-of-sample accuracy. This pattern of results suggests that the out-of-sample performance we report here is not the maximum that can be achieved with the present approach; more training data (more judgment targets) and/or lower variance models could further improve out-of-sample accuracy.

Alternative embedding techniques and corpora

In our primary analyses, we used pre-trained word2vec embeddings because these contained embeddings for names, brands, and other common multi-word expressions for which we elicited judgments. However, to show the generality of our approach, we tested a number of other pre-trained embedding models, varying both in training algorithm and in training corpus. SI section Embedding Techniques has more details of the models we used, but in brief, we tested GloVe (Pennington, Socher, & Manning, 2014), fastText (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018), Paragram-SL999 (Wieting, Bansal, Gimpel, Livescu, & Roth, 2015), post-specialised GloVe embeddings (Ponti, Vulic, Glavas, Mrksic, & Korhonen, 2018), and ELMO (Peters et al., 2018) and BERT (Devlin, Chang, Lee, & Toutanova, 2018) embeddings obtained in two different ways. The first approach computes a vector for each judgment target by putting it in a "sentence" with the domain name, e.g., "food bass" (e.g., "ELMO, in context" in **Figure 3C**). This extracts a vector for the sense of the word relevant to the current domain (i.e., "bass" as in the fish and not the musical instrument). The second approach computes a vector without any such contextualizing (e.g., "ELMO, no context" in **Figure 3C**). Because many of our brands and almost all of our people were multi-word expressions which were overwhelmingly missing from these additional embedding spaces, we only compare word embedding models on traits, risk sources, foods, occupations, and consumer goods.[1] Finally, for each embedding type, we performed our cross-validation procedure only for ridge models.

Figure 3C has out-of-sample correlations for every judgment dimension on which we compared the various embedding models. As can be seen, performance is quite good across many embedding models and judgment dimensions, but there is substantial variation across embedding types. It also seems that the prediction of masculinity and femininity judgments in particular benefit from embeddings specialized for word-word similarity: compare Paragram-SL999 to word2vec, and Glove-Postspec to Glove. This may be because our traits have a number of antonym pairs among them (e.g., loyal vs. disloyal). Whereas typical word embedding techniques (word2vec, glove) often assign nearby vectors to such antonym pairs, antonyms tend to be further apart from one another in Paragram-SL999 and GloVe Post-spec space. However, Paragram-SL999 and Glove Postspec perform somewhat worse on the rest of the judgment dimensions, leading to their poorer overall accuracy. This leaves an ambiguous picture regarding the importance of using word embeddings that respect word-word similarity relations to model at least the present kinds of judgments (cf. the utility of similarity-specialised embeddings in other downstream NLP tasks like sentiment analysis [Wieting et al., 2015] and lexical text simplification [Ponti et al., 2018]). Aside from this, it is a bit difficult to interpret differences between the embedding models, given the conflation we noted before between training algorithm and training corpus. Our main point, however, is that our approach is robust to a range of training algorithms and corpora.

Psychological Substrates of Judgment

The ridge regression approach used in most of the above tests involves learning a (regularized) linear mapping from the semantic space to the judgment dimension. The best-fit weights for this mapping have the same dimensionality as the semantic space, and can thus be seen as representing a vector in this space. Judgment items whose vectors project strongly onto the weight vector (typically judgment items whose vectors are highly similar to the weight vector) will be predicted to have the highest judgment ratings. Given this interpretation, we can ask what other objects and concepts (that may not necessarily be judgment targets themselves) project strongly onto the weight vector. Intuitively, these would be the objects and concepts that are most related to the judgment, and may correspond to the judgment-relevant qualities that people evaluate when generating their responses. Uncovering these words and concepts would shed light on the psychological substrates of the judgment in consideration.

We used two methods to uncover these psychological substrates. Our first method involved obtaining word embedding representations for the 5,000 most common words in English that were not also judgment targets, and passing these embeddings through our trained ridge regression mapping to determine the associations of these words with our 14 judgment dimensions. We then computed the difference between a word's predicted association with one dimension (e.g., masculinity) and its predicted association with the complementary dimension (e.g., femininity), to find the words most strongly associated with one dimension relative to the other. Our second method used a similar approach, applied to word lexicons corresponding to 256 core psychological constructs from the General Inquirer (GI) and Linguistic Inquiry and Word Count (LIWC) dictionaries (Pennebaker, Boyd, Jordan, & Blackburn, 2015; Stone, Dunphy, & Smith, 1966). In the SI we present additional details of these methods, as well as the results of these methods (see Figures S5 and S6 for summary visualizations, and Tables S3 and S4 for exhaustive reuslts). These results show, for example, that work-related interpersonal concerns are most associated with masculine traits whereas home and family concerns are most associated with feminine traits; passivity and submission are associated with high-warmth/low-competence individuals and strength and hostility are associated with high-competence/low-warmth individuals; natural objects are highly associated with nutrition; moral values and values of love and friendship are associated with sincere brands whereas values of power and respect are associated with exciting brands; and need-related motivations are associated with hedonic goods whereas means-related motivations are associated with utilitarian goods. These results conform with our intuition and with prior empirical work (Cuddy, Fiske, & Glick, 2008; Heilman, 2012; Rozin et al., 2004), despite not explicitly eliciting judgments about psychological constructs such as work-related or home-related interpersonal concerns and means-related or need-related motivations. Of course it is also possible to use these results to generate novel

behavioral hypotheses, in a bottom up manner, which can be tested in subsequent empirical work.

Discussion

Despite the ubiquity of human judgment, until now we have had limited ability to predict arbitrary human judgments of objects and concepts, as capturing the rich knowledge used to make predictions has been difficult or impossible. Here we demonstrated in a pre-registered study that word embeddings – vector representations for words and concepts based on statistics of language use – proxy this knowledge and can predict 14 diverse judgments across the behavioral sciences with a high degree of accuracy, at both a group- and individual-level. We also showed that the learned mapping from word embeddings to judgments can also be used to explore the conceptual underpinnings of judgments, by mapping non-judgment target entities onto the judgment dimension.

We view the present approach as a modern extension to classical psychometric approaches used to uncover the underlying representations used for making judgments (McRae et al., 2005; Shepard, 1980; Slovic, 1987). However, the present approach offers several advantages over classical techniques. First, the only human data that our approach requires is a (relatively) small number of judgment ratings to train a predictive model. Once a satisfactory model has been trained, no new human psychometric data is required to predict judgments for new entities. Second, word embeddings provide a single representational space to predict and understand behavior across many important domains, judgment target entities (objects, people, organizations, social relations, traits), and judgment dimensions. We believe that using a single representational space to model judgments relevant to different disciplines can provide cohesion to the behavioral sciences that traditionally rely on different methods and data. Third, word embeddings – owing to their high dimensionality – capture more knowledge about judgment targets than can realistically be collected from human participants, especially when the relevant knowledge used to make a particular judgment is not already theoretically well-understood and thus surveyed from human participants. Capturing a great degree of knowledge leads to the high predictive accuracy we have achieved here, which we suggest may be high enough for applications in downstream behavioral sciences and technologies.

We conclude by considering our work in the context of a recent meta-theoretical conversation in the social and behavioral sciences. As has been argued previously (Hofman, Sharma, & Watts, 2017; Yarkoni & Westfall, 2017), social, cognitive, and behavioral scientists have traditionally been focused on interpretable, explanatory models, to the detriment of developing models that make accurate out-of-sample predictions. Of course, this is undesirable to the extent that we think a good model requires external validity – having statistically significant, interpretable model coefficients is ultimately of limited use if a model can't predict new behavior with any accuracy. While some have claimed that behavioral scientists may sometimes need to choose between predictive or interpretable models (Yarkoni & Westfall, 2017), our models of judgment achieved unprecedented predictive accuracy, and substantial progress towards explanatory insights into judgments (sections *Amount of Information Required for Prediction* and *Psychological Substrates of Judgment*). We thus believe word embeddings are an excellent step towards fully predictive and interpretable models of the thousands of diverse and consequential judgments people make every day.

Methods

We recruited 354 participants (mean age = 31.89 years, 46.19% female) through Prolific Academic. We limited our data collection to participants who were from the U.S. and had an approval rate above 80%. Participants were only allowed to participate once, and they were paid $4.40 each.

Using a between-subjects design, we randomly assigned each participant to one of the seven judgment domains – brands (N = 54), consumer goods (N = 51), traits (N = 46), foods (N = 55), occupations (N = 49), risk sources (N = 49), people (N = 51). After being randomly assigned to one judgment domain, participants were instructed to rate 200 items on two dimensions from −100 (e.g. not at all significant) to 100 (e.g. extremely significant), one item at a time (see Figure S1 in the SI for scatterplots of judgments for all seven judgment domains). They were asked to check "not applicable" if they were not familiar with a particular item. There was a 30-second break after every 50 items, although participants were allowed to proceed without breaks if they so chose. Participants' age and gender information were collected at the end of the study. The mean completion time was 38 minutes; the median was 33 minutes.

In line with our pre-registered analysis plan, we excluded participants who took more than 1 hour to complete the study or indicated being unfamiliar with more than 25% of the items. Accordingly, one person was dropped from the occupation condition, one from risk, 11 from brands, 16 from people, and 36 from foods. An item was excluded if fewer than 25% of participants indicated being familiar with it. One brand, three people, and 38 foods were accordingly dropped. Retaining all items and participants does not affect our main result substantially: for a ridge regression with λ set to 10, the mean Pearson correlation coefficient between predicted and actual judgments is .763 vs .769 under our pre-registered exclusion criteria. Taste and nutrition have the largest drop in performance, from $r = .66$ to $r = .59$ and $r = .83$ to $r = .79$, respectively, when retaining all items and participants.

Detailed instructions, survey questions and items, and exclusion criteria can be found at https://osf.io/t7qyb/.

We adopt our baseline/similarity-based approach from Grand et al. (2018). This method works as follows: First, we select words reflective of high and low ends of some judgment dimension. For example, the occupation significance dimension was represented by the words significant, meaningful, important and insignificant, meaningless, unimportant, pointless. Where possible,

we chose words used in previous literature to define the dimensions (see SI for these words). Then, for each judgment dimension, the average pairwise vector difference between each possible pair of high and low words is computed to obtain a single vector *d* representing that dimension. Last, to obtain a score for a judgment target entity on that dimension, we compute the dot product between the target entity's embedding x_i and the dimension embedding, $d * x_i$. As stated in Results, this method essentially computes the similarity of a judgment target to words high relative to words low along the dimension of interest.

Our mapping approach trains various supervised learning algorithms to predict judgments from word embeddings. More formally, we are training a model to predict a numerical rating y_i for a judgment target *i*, from its 300 dimensional word embedding x_i, where x_{ij} is the value of the judgment target on dimension *j* of its embedding. Given that we have fewer judgment targets (< = 200) than independent variables (300), we chose three regularized regression techniques: lasso, ridge, and support vector machines. We also included k-nearest neighbors (KNN) regression since it was used in work on predicting risk perceptions with word embeddings (Bhatia, 2019). We describe our secondary models (lasso, SVR, KNN) more in the SI, and instead focus here on our primary model, ridge regression. Ridge regression entails learning a vector of coefficients β_j for the 300 dimensions of our embeddings, such that the following loss expression is minimized:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

The left term of the loss expression is merely ordinary least squares regression. The additional sum on the right is a "penalty term" which pushes coefficients towards 0, depending on the strength of the regularization parameter, λ. We searched for the optimal penalty strength by a cross-validation procedure described in the SI, and found λ = 10 yielded the best out-of-sample r-squared and root mean squared error. Ridge with λ = 10 is thus the model we use throughout the paper. (We performed a similar search for our other models' optimal hyperparameters). All our supervised models were implemented in the Scikit-Learn machine learning library for Python (Pedregosa et al., 2011).

Notice that ridge (and lasso, or any other linear mapping) is like our similarity-based approach in that both produce a judgment rating by computing the dot product between a word embedding and a vector representing a judgment dimension. The difference between the approaches is that the similarity approach stipulates the vector for the judgment dimension, while the mapping approach learns the best vector given some true human judgments. While the similarity approach can score a judgment target on some dimension without any human data, these scores correlate with actual human judgments far worse than do predictions generated by the mapping approach. Whether one prefers accuracy or the lack of need for training data is likely to depend on one's subjective preferences or goals as a modeler (Hofman et al., 2017). Applied behavioral scientists, at least, likely want accurate predictions of human judgments so that they can predict important downstream cognition and behavior (see "Example Applications" of **Table 1**).

Data Accessibility Statement

Data and code producing the reported analyses can be found at https://github.com/drussellmrichie/embeddings_to_judgments.

Note

[1] In principle, one could obtain, e.g., GloVe vectors for people's names and other MWE's just as in word2vec. All that is required is tokenizing the desired MWE's as individual "words" in the corpus before computing the global co-occurrence matrix for GloVe.

Acknowledgements

Thanks to members of the Computational Behavioral Science Lab and the Integrative Language Science and Technology seminar for their feedback on this work.

Funding Information

Funding by National Science Foundation grant SES-1626825 and the Alfred P. Sloan Foundation was provided to the third author.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

R.R., W.Z., and S.B. designed the study, R.R. and W.Z. collected data, R.R. and S.B. performed analyses, R.R., W.Z., and S.B. wrote the paper.

References

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research*, 347–356. DOI: https://doi.org/10.1177/002224379703400304

Batra, R., & **Ahtola, O.** (1990). Sources of the hedonic and utilitarian measuring attitudes consumer. *Consumer Attitudes*, *2*(2), 159–170. DOI: https://doi.org/10.1007/BF00436035

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*(2), 155–162. DOI: https://doi.org/10.1037/h0036215

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20. DOI: https://doi.org/10.1037/rev0000047

Bhatia, S. (2018). Semantic processes in preferential decision making. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. DOI: https://doi.org/10.1037/xlm0000618

Bhatia, S. (2019). Predicting risk perception: New insights from data science. *Management Science*, *65*(8): 3800–3823. DOI: https://doi.org/10.1287/mnsc.2018.3121

Bhatia, S., & **Olivola.** (2018). Computational brand perception. Working paper.

Bruni, E., Tran, N.-K., & **Baroni, M.** (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47. DOI: https://doi.org/10.1613/jair.4135

Caliskan, A., Bryson, J. J., & **Narayanan, A.** (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. DOI: https://doi.org/10.1126/science.aal4230

Cuddy, A. J., Fiske, S. T., & **Glick, P.** (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, *40*, 61–149. DOI: https://doi.org/10.1016/S0065-2601(07)00002-0

Cuddy, A. J., Fiske, S. T., Glick, P., & **Xu, J.** (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. DOI: https://doi.org/10.1037/0022-3514.82.6.878

Devlin, J., Chang, M.-W., Lee, K., & **Toutanova, K.** (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Garg, N., Schiebinger, L., Jurafsky, D., & **Zou, J.** (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. DOI: https://doi.org/10.1073/pnas.1720347115

Grand, G., Blank, I. A., Pereira, F., & **Fedorenko, E.** (2018). Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv preprint arXiv:1802.01241*.

Hackman, J. R., & **Oldham, G. R.** (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, *16*(2), 250–279. DOI: https://doi.org/10.1016/0030-5073(76)90016-7

Healey, M. K., & **Kahana, M. J.** (2016). A four-component model of age-related memory change. *Psychological Review*, *123*(1), 23–69. DOI: https://doi.org/10.1037/rev0000015

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in organizational Behavior*, *32*, 113–135. DOI: https://doi.org/10.1016/j.riob.2012.11.003

Hill, F., Reichart, R., & **Korhonen, A.** (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695. DOI: https://doi.org/10.1162/COLI_a_00237

Hills, T. T., Jones, M. N., & **Todd, P. M.** (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. DOI: https://doi.org/10.1037/a0027373

Hofman, J. M., Sharma, A., & **Watts, D. J.** (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488. DOI: https://doi.org/10.1126/science.aal3856

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & **Jacobs, A. M.** (2018). Simple co-occurrence statistics reproducibly predict association ratings. *Cognitive Science*, *42*(7), 2287–2312. DOI: https://doi.org/10.1111/cogs.12662

Hollis, G., Westbury, C., & **Lefsrud, L.** (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, *70*(8), 1603–1619. DOI: https://doi.org/10.1080/17470218.2016.1195417

Jones, M. N., Kintsch, W., & **Mewhort, D. J.** (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. DOI: https://doi.org/10.1016/j.jml.2006.07.003

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171. DOI: https://doi.org/10.1146/annurev-linguistics-030514-125254

Levy, J., Bullinaria, J., & **McCormick, S.** (2017). Semantic vector evaluation and human performance on a new vocabulary MCQ test. In *Proceedings of the 39th annual conference of the cognitive science society* (pp. 2549–2554).

Mandera, P., Keuleers, E., & **Brysbaert, M.** (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. DOI: https://doi.org/10.1016/j.jml.2016.04.001

McRae, K., Cree, G. S., Seidenberg, M. S., & **McNorgan, C.** (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. DOI: https://doi.org/10.3758/BF03192726

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & **Joulin, A.** (2018). Advances in pre-training distributed word representations. In *Proceedings of the international conference on language resources and evaluation (lrec 2018).*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & **Dean, J.** (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & **Blackburn, K.** (2015). *The development and psychometric properties of liwc2015 (Tech. Rep.).*

Pennington, J., Socher, R., & **Manning, C.** (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). DOI: https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & **Zettlemoyer, L.** (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*. DOI: https://doi.org/10.18653/v1/N18-1202

Pilehvar, M. T., & **Camacho-Collados, J.** (2018). Wic: 10,000 example pairs for evaluating context-sensitive representations. *CoRR*, *abs/1808.09121*. Retrieved from http://arxiv.org/abs/1808.09121

Ponti, E. M., Vulic, I., Glavas, G., Mrksic, N., & **Korhonen, A.** (2018). Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. *CoRR*, *abs/1809.04163*. Retrieved from http://arxiv.org/abs/1809.04163. DOI: https://doi.org/10.18653/v1/D18-1026

Raghunathan, R., Naylor, R. W., & **Hoyer, W. D.** (2006). The unhealthy= tasty intuition and its effects on taste inferences, enjoyment, and choice of food products. *Journal of Marketing*, *70*(4), 170–184. DOI: https://doi.org/10.1509/jmkg.70.4.170

Rosenberg, S., Nelson, C., & **Vivekananthan, P.** (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283–294. DOI: https://doi.org/10.1037/h0026086

Rozin, P., Spranca, M., Krieger, Z., Neuhaus, R., Surillo, D., Swerdlin, A., & **Wood, K.** (2004). Preference for natural: instrumental and ideational/moral motivations, and the contrast between foods and medicines. *Appetite*, *43*(2), 147–154. DOI: https://doi.org/10.1016/j.appet.2004.03.005

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398. DOI: https://doi.org/10.1126/science.210.4468.390

Slovic, P. (1987). Perception of risk. *Science*, *236*(4799), 280–285. DOI: https://doi.org/10.1126/science.3563507

Stone, P. J., Dunphy, D. C., & **Smith, M. S.** (1966). The general inquirer: A computer approach to content analysis.

Turney, P. D., & **Pantel, P.** (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. DOI: https://doi.org/10.1613/jair.2934

Wieting, J., Bansal, M., Gimpel, K., Livescu, K., & **Roth, D.** (2015). From paraphrase database to compositional paraphrase model and back. *arXiv preprint arXiv:1506.03487*. DOI: https://doi.org/10.1162/tacl_a_00143

Yarkoni, T., & **Westfall, J.** (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. DOI: https://doi.org/10.1177/1745691617693393

How to cite this article: Richie, R., Zou, W., & Bhatia, S. (2019). Predicting High-Level Human Judgment Across Diverse Behavioral Domains. *Collabra: Psychology,* 5(1): 50. DOI: https://doi.org/10.1525/collabra.282

Senior Editor: Simine Vazire

Editor: Simine Vazire

Submitted: 19 August 2019 **Accepted:** 18 September 2019 **Published:** 23 October 2019