Human word similarity judgments replication by word-embedding

QUEENSLAND UNIVERSITY OF TECHNOLOGY

Roberto Carlos da Silva Junior
Student ID: n10374647
Master of Information technology

ABSTRACT

Judgment is a natural part of human behavior that lies in every sphere of life. An example is in linguistic judgment, in which people judge the meanings of words and compare them with others with the same context. Behavioral scientists have struggled to find a model that predictive and explanatory models to quantify the semantic similarity between lexical items based on their contexts as these models demand a high volume of data, which is not easy to be collected as most depend on asking people rating. Thus, Word embedding, a model in the computer area, permits the outperform N-gram models through artificial intelligence, vector representations for millions of objects and concepts based on word use statistics in large language corpora that can accurately replicate human judgments (Mikolov et. al., 2013). Some literature proposes using word embedding to replicate human judgments in linguistic such as the paper: Efficient Estimation of Word Representations in Vector Space that presents a technique to training large-scale models in accuracy at much lower computational cost (Mikolov et. al., 2013). Furthermore, Extracting semantic representations from word co-occurrence statistics: A computational study that presents a systematic exploration of the principal computational possibilities for formulating and validating the meaning of words from word co-occurrence statistics (Bullinaria & Levy, 2006) were analyzed. However, for this paper, the technique used is demonstrated in Semantic representations extracted from large language corpora to predict high-level human judgment in seven diverse behavioral domains. Which presents a technique developed by the Department of Psychology and Marketing of the University of Pennsylvania to applies word embedding to simulate human judgments in seven different domains of linguistic and also evaluates how well this model replicates these judgments. (Echizen'ya et. al., 2019).

INTRODUCTION

In linguistic and semantic tasks, people are always judging the meanings of words and comparing these words with others with the same context (Nagata, 1987). Thus, behavioral scientists must develop predictive and explanatory models to quantify the semantic similarity between lexical items based on their contexts (Mikolov et al., 2013). However, executing these methods has been a challenge as most of those, such as the psychometric, face serious challenges such as the cost to collect the information needed to apply its approaches, as it depends on asking participants to rate target entities (Kovacs & Kleinbaum, 2019). What makes it almost impossible to apply its approaches when it is needed to quantify representations for judgment across hundreds of thousands of behavioral domains (McRae et. al., 2005).

Thus, a technique is needed to classify outcomes for many objects to model judgments. What is found in Hyperspace Analogue to Language (HAL), which as *psychometric* approaches, depending on the similarity of objects to discover representations of entities. However, these are not detailed participant similarity ratings, so word-based vocabularies in the form of vectors, in which an individual word turned into numeric-valued, lie within a vector containing other numeric-valued words with the same context, and this vector is mapped with other vectors-contexts in a way that resembles a neural network. In which the vectors themselves are typically far more abundant than *psychometric* techniques (Naili, 2017).

Word-embedding is the cut edge in HAL that revealed that it is possible to replicate complex judgments with a high degree of precision (Senel et. al., 2018). What to affirm so, human inter-rater reliability metrics have been using to score how much homogeneity or consensus exists in the ratings given by various judges. However, throughout a study (Richie et al., 2019), it reached high predictive accuracy using ridge regression with hyperparameters, which affirms to present better outcomes than inter-rater reliability metrics (Richie, 2019).

Therefore, this paper explores the technique presented by Richie et al. (2019) to estimate the prediction accuracy of an extensive model with judgment samples. This prediction accuracy of the same model will also be measured by human inter-rater. The idea is to compare the measurement from both methods to figure out how well the model proposed measures the model's prediction accuracy comparing its predictions by human inter-rater reliability metrics, as human inter-rater reliability is considered an upper bound in this kind of prediction Richie et al., (2019). What will result in a literature review that proves that word-embedding replicates human word judgments and does it in a high degree of accuracy. However, before the results be presented is crucial to understand the key techniques of both methods.

FROM NEURAL NETWORKS TO SKIP-GRAM

The neural networks are inspired by the brain's functioning, the ability to recognize patterns and learn from mistakes and successes. These models use artificial neurons, which are a very simplified abstraction of a biological neuron. A neural network can be seen as a system of interconnected neurons that process information in response to external inputs. Generally, the neural networks models are organized into three layers: a layer as an input, a hidden layer, and a layer as an output. The Figure 1 is an example of the structure of a neural network, in which each node represents an artificial neuron, and the arrows represent the connection of the output of one neuron with the input of another. The input of each neuron is a weighted sum, and this value is used as an input of an activation function.

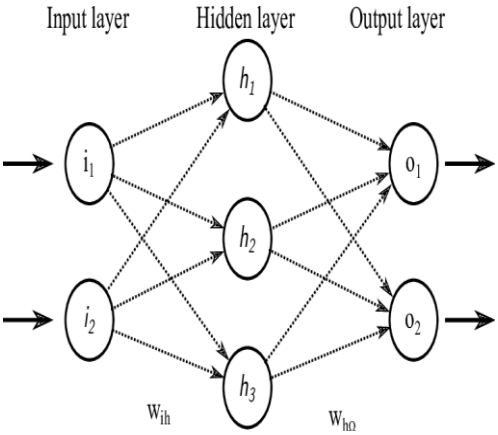

Figure 1: Example for an artificial neural network with two input neurons, two hidden neurons and two output neurons, connected by synapses (Jahr et. al., 2015).

Neural networks need much data for training, and they also need data for testing (Jahr et. al., 2015). The training data represent data used to create the model by the neural network (called pre-trained model). These data usually represent about 70% of the total data. Whereas the test model represents the model after its creation, simulating real predictions that the model will make, thus allowing the real performance to be verified. These data usually represent about 30% of the total data. Nowadays, most of the NLP problems involving processing semantic and syntactic and linguistic varieties are solved by neural networks.

Natural language processing, or just NLP, is a set of computational techniques that deal with human language problems, both written and spoken. Its main challenges encompass the understanding and extraction of meaning from language. The NLP has essential applications such as information extraction, spelling correction, information retrieval, voice recognition, and automatic.

The works related to the NLP appeared at the end 1940s, with a focus on machine translation. One of the first research in the field was an automatic translation from Russian to English in a rudimentary and limited experiment, which was applied in the demonstration IBM-Georgetown in 1954 (Karen, 1994). In NLP applications, we cannot directly feed their models by textual data. It needs to convert these text data into numerical form, then fed these models for further processing.

Word Embedding converts textual data into numeric data in some way, such as vector representation. As an example of word embedding in real world (In this example the vector will represent people and not word), it is simulated a Big Five Personality Traits test, which is the most widely accepted personality theory held by psychologists today. This test asks an interviewer a question giving a determined scale, then scores on several axes, one of which is introversion/extroversion. For instance, giving the examples: Openness to experience 79 out of 100, Agreeableness 75 out of 100, Conscientiousness 42 out of 100, Negative emotionality 50 out of 100 and Extraversion 58 out of 100. How introverted/extraverted are you? (where 0 is the most introverted and 100 is the most extroverted).

To follow the example, a person called Jay scored 38/100 as introversion/extraversion score and so it can be traced in this way as in Figure 2. Which converted to the range -1 to 1 will look like as in Figure 3.

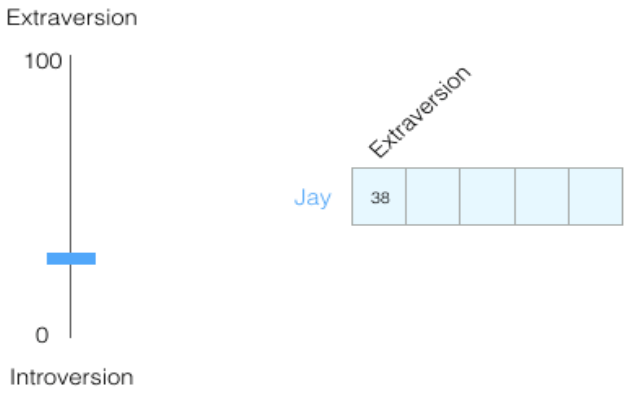

Figure 2 – A person called Jay scored 38/100 as introversion/extraversion score (Jay, 2018).

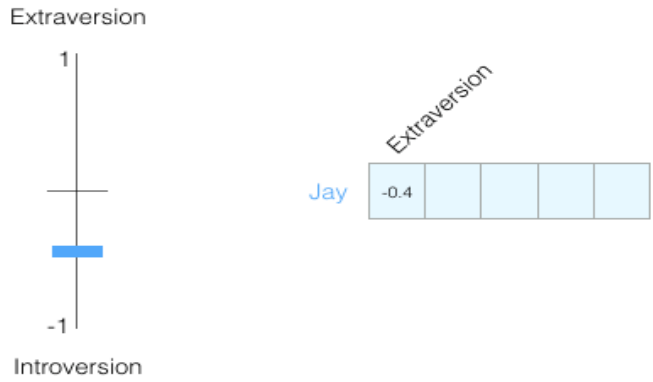

Figure 3 – Jay scores 38/100 is converted to the range -1 to 1 (Jay, 2018).

However, it is hard to know a person just by this introversion/extroversion information, and so another dimension is added, which is the score for another feature of the test.

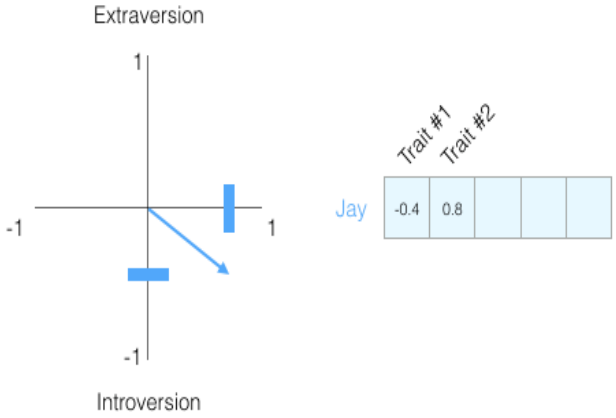

Figure 4 – Another dimension is added to Jay vector (Jay, 2018).

The two dimensions can be represented as points on a graph. It may be said that this vector represents the personality of a person. The practicality of such a representation comes when it is wished to compare two other people to Jay. For instance, Jay was hit by a bus, and he cannot work, making necessary its replacement by someone with a similar personality. In the figure below, which of the two people is most like Jay?

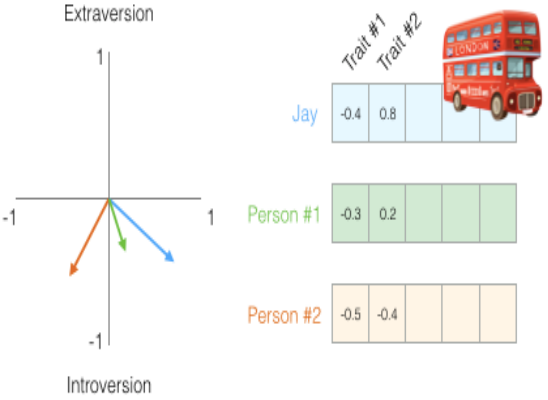

Figure 5 – Representation of Person #1 and Person #2 as vector (Jay, 2018).

A common way to calculate a vector similarity score is cosine similarity. According to it, Person # 1 is more like Jay in personality as their vectors point in the same direction, having a higher cosine similarity value.

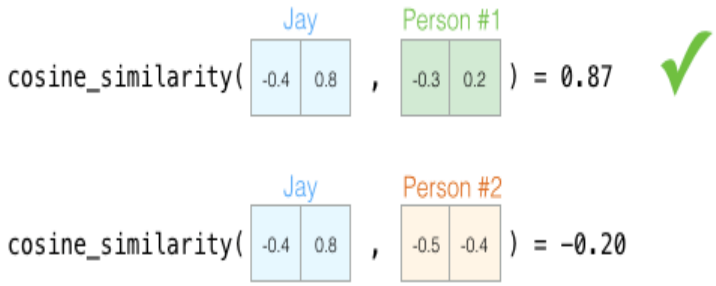

Figure 6 – Cosine similarity for vectors Jay, Person #1, and Person #2 vector (Jay, 2018).

Due to the Word Embedding studies' advance in recent years, many Word Embedding algorithms have come out, such as Word2vec (Jay, 2018).

Word2Vec is a statistical method for efficiently learning a standalone word embedding from a text corpus. It was developed in 2013 as a response to make the neural-network-based training of the embedding more efficient and since then has become the de facto standard for developing pre-trained word embedding.

Word2Vec offers two architectures as an option: continuous bag-of-words (CBOW) and skip-gram. In the CBOW model, the input is a context, and the output is a target word. Whereas, in the skip-gram a word is the input to predict a target context. Both architecture are illustrated in Figure 7.

As an example of both architectures, following the example above, fill in the blank: **Jay was hit by a** ____.

The phrase given above is five words before the blank word. Moreover, the word “bus” is easy to be guessed if the example above was followed. However, if given one more piece of information – a word after the blank- would that change the answer? **Jay was hit by a** ____ **bus**.

This completely changes what should go in the blank. The word red is now the most likely to go into the blank. From what was said above, the words both before and after a specific word carry informational value. It turns out that accounting for both directions (words to the left and to the right of the word to be guessed) leads to better word embeddings. Thus the way the model has been training can be adjusted to account for this. Instead of only looking at two words before the target word, it can also look at two words after it. For instance, **Jay was hit by a** ____ **bus in…** [by | a | red | bus | in]. If this is done, the dataset virtually built and training the model against will look like this: Input = by / Input = a / Input = bus / Input = in / output = red. That is how CBOW works.

However, instead of guessing a word based on its context (the words before and after it), skip-gram tries to guess neighboring words using the current word. For instance, from the phrase **Jay was hit by a red bus in…** Skip-gram architecture would create four separate samples in its training dataset, one per output**. Such as input = Red and output = by, input = Red and output = a, input = Red and output = bus, and input = Red and output = in** (Jay, 2018).

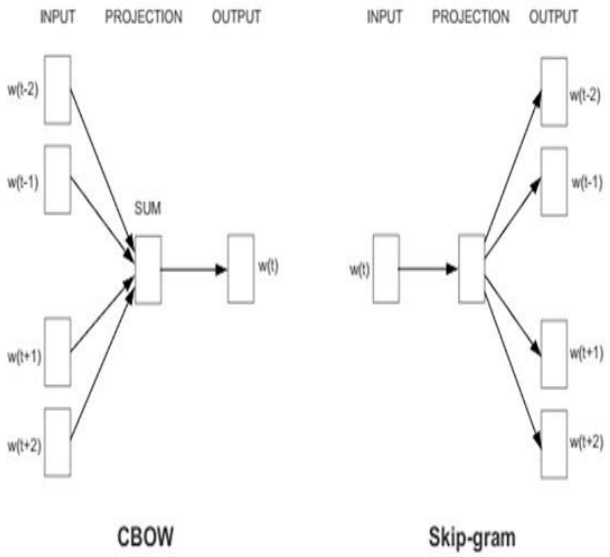

Figure 7 - The two Word2vec architectures: continuous bag-of-words (CBOW) and skip-gram (Karen, 1994).

In order to explain how Word Embedding techniques could solve a problem in real world, this paper demonstrated a Big Five Personality Traits test simulation, which finished showing the result of similarity score of vectors by the cosine similarity.

In Vector Space Model, Cosine is largely used to measure the similarity between two vectors as in the example given, being very efficient when only the non-zero dimensions need to be considered, and so, it has been used to solve diverse text mining problems. However, cosine similarity is not efficient when used to higher values features and does not care much about how many features two vectors share (Li & Han, 2013). Another way to measure the similarity between vectors is from regression.

FROM REGRESSION ANALYSIS TO LEAVE-ONE-OUT CROSS-VALIDATION

Regression is a data mining technique used to predict a range of numeric values, given a particular dataset. For instance, a company based on current economic conditions wants its growth in sales to be estimated. The recent company data indicates that the growth in sales is around two and a half times the economy's growth. From this insight, the future company sales can be predicted based on current & past information. Although the example given is nothing related to sematic and linguist tasks, it is an easy way to understand the concept of regression analysis.

There are various kinds of regression techniques available to make predictions, such as the Linear Regression, which is a model that aims to summarize the relationship between two or more variables through a straight line that is defined by the equation $Y = a + b \times X$, and thus use the result of the function of that line to estimate values when knowing the variables that affect it. Which the variable to be discovered is the Y of the function, called the dependent variable. Whereas, the variables (X) that influence the Y, it is called independent variables, and y-intersect of the line and b is its slope, for instance, a scatter plot and the corresponding regression line and regression equation for the relationship between the dependent variable body weight (kg), and the independent variable height (m) is illustrated in Figure 8.

The Loss function is essential in Linear Regression because it determines the error between the outcome (Y) and the given target value. In other words, the loss function expresses how far off the mark the output is. When the independent variables (Y) are highly correlated, that means their variances are large. It is used a technique called Ridge Regression, which adds degrees of bias to the regression estimates to stability their Y by adding additional costs called hyperparameters to the loss function (Schneider et. al., 2010).

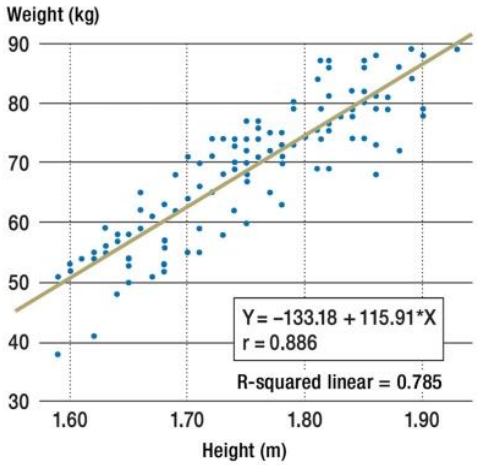

Figure 8 - Line and regression equation for the relationship between the dependent variable body weight (kg) and the independent variable height (m) (Schneider et. al., 2010).

As said already mentioned, Regression is a data mining technique used to predict a range of numeric values, given a particular dataset. This dataset is composed of different variables, in which statistical measurement technics called Correlation coefficients are used to calculate the strength of the relationship between these variables. There are several

types of Correlation coefficients. However, the most used is the Pearson product-moment correlation coefficient, also known as r, R, or *Pearson's* r, which measures the strength and direction of the linear relationship between two variables divided by the product of their standard deviations (Waldmann, 2019).

In order to verify whether a model is performing as expected, it is used a procedure called Cross-validation. This procedure evaluates models on a limited data sample by estimating their prediction error, and it has a single parameter called k that refers to the number of groups a given data sample is to be split. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in reference to the model, such as k=10 becoming 10-fold cross-validation. Leave-one-out cross-validation (LOOCV) is a kind of K-fold cross-validation, in which it takes a single data of the set as validation (P) and K is equal to the number of data points in the set (N). That means it verifies all the data except for the one chosen as the validator (P) (Tzu-Tsung, 2015). The homogeneity of the raters can also be measure, for this exist a technique called Inter-rater reliability.

INTER-RATER RELIABILITY

Inter-rater reliability scores the degree of agreement among raters; That is, it scores how much homogeneity or consensus exists in the ratings given by various judges. Inter-rater reliability can be evaluated by using a number of different statistics. Some of the more common statistics include percentage agreement, kappa, product-moment correlation, and intraclass correlation coefficient. Examples of inter-rater reliability are the check the homogeneity of the variables body weight (kg) and height (m) given in Figure 8 example (Lange, 2011).

LITERATURE REVIEW METHODOLOGY

In order to prove that word-embedding replicates human similarity judgment, a research review occurred in detail in the literature proposed, Predicting High-Level Human Judgment Across Diverse Behavioral Domains by Richie et al., 2019, which ensure a better understanding of the problem and enable a full assessment of the current scenario of the use of ridge regression to evaluate the accuracy prediction of judgments. Bellow is described in detail how the literature collected data and apply the proposed technique to measure the accuracy of the judgment.

SUMMARY OF DATA

The **pre-trained model** is a very large dataset of Google News articles containing 300-dimensional vectors for 3 million judgment words and short phrases trained by Word2vec, whereas, the **test model** is a sample of data that involved 140,000 participant judgments. Each participant answered how much they judged each of the seven behavioral domains (Giving two Judgment dimensions' examples per behavioral domain) on a scale of -100 to 100. For instance, for the behavioral domain Brand, the question was, "How sincere or exciting (Judgment dimensions) do you feel the following **brand** Home Depot (entities) is? A brand is sincere if it is down-to-earth, honest, wholesome, and cheerful, and a brand is exciting if it is daring, spirited, imaginative, and up-to-date. If you are unfamiliar with this brand, check the box at the right. ". The question of all behavioral domain is described bellow. Whereas a summary of judgment dimensions and items we consider, along with example fields and applications and relevant references, is provided in Table 1. And a summary containing the two Highest rated items across domains in shows in Table 2.

Question for behavioral domain brand.

How {sincere/exciting} do you feel the following brand is? A brand is sincere if it is down-to-earth, honest, wholesome, and cheerful, and a brand is exciting if it is daring, spirited, imaginative, and up-to-date. If you are unfamiliar with this brand, check the box at the right.

Question for behavioral domain object.

How much {hedonic|utilitarian value) do you feel the following object has? An object is hedonic if it is fun, exciting, delightful, thrilling, and/or enjoyable. An object is utilitarian if is effective, helpful, functional, necessary, and/or practical. If you are unfamiliar with this object check the box at the right.

Question for behavioral domain trait.

How {masculine/feminine} do you feel the following trait is? For our purposes, a trait is masculine if society generally desires it in men, and a trait is feminine if society generally desires it in women. Note: your rating should not reflect how desirable you find these traits in men/women, but rather how desirable society in general finds these traits in men/women. If you are unfamiliar with this trait, check the box at the right.

Question for behavioral domain trait.

How {tasty/nutritious} do you feel the following food is? By 'tasty', we simply mean that the food tastes good; by 'nutritious', we mean that the food is generally 'good for you'. If you are unfamiliar with this food, check the box at the right.

Question for behavioral domain trait.

How much {significance/autonomy} do you feel the following job has? A job is significant if it has a substantial impact on the lives or work of other people, whether in the immediate organization or in society in general. A job provides autonomy if the job provides substantial freedom, independence, and discretion to the individual in determining where, when, and how the work is done.

Question for behavioral domain trait.

How {dread-inducing/unknowable} do you feel the following potential risk source is? A potential source of risk is 'dread'-inducing if it is uncontrollable, induces feelings of dread, its risks are globally catastrophic, and it has fatal consequences. A potential source of risk is 'unknowable' if the risk source is not observable, is unknown to those exposed to it, has delayed effects, is new, and is unknown to science. If you are unfamiliar with this potential risk source, check the box at the right.

Question for behavioral domain trait.

How {warm|competent} do you feel the following person is? For our purposes, someone is warm if they are generous, helpful, sincere, tolerant and warm, and someone is competent if they are efficient, foresighted, creative, competent, and intelligent. If you are unfamiliar with this person, check the box at the right.

Table 1 – Two highest items across the behavioural domains.

Behavioral domain		
Brand	Good	Trait

Judgment dimensions	**Sincere**	Lego	57.311111	**Hedonic**	chocolate	80.673913	**Masculine**	tough	79.289474
		Costco	60.184211		artwork	80.934783		powerful	79.707317
	Exciting	Lego	57.00000	**Utilitarian**	trashcan	91.847826	**Feminine**	loving	73.461538
		NetFlix	58.72093		diaper	91.891304		attractive	74.128205
	Food			Occupation			Risk		
	Tasty	Cheese	83.307692	**Significance**	Neurosurgeon	82.883721	**Dread-inducing**	war	70.775000
		Sweets	84.098039		Surgeon	83.302326		cancer	74.125000
	Nutritious	carrots	84.961538	**Autonomy**	Hacker	74.441860	**Unknowable**	spies	47.743590
		spinach	88.192308		Thief	76.487805		cancer	50.775000
	People								
	Warm	Harriet Tubman	67.333333						
		Mother Teresa	84.098039						
	Competent	Leonardo Da Vinci	85.600000						
		Wolfgang Amadeus Mozart	85.744186						

Table 2: 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 (Richie et al., 2019, p. 4).

Behavioral domains	Judgment dimensions	Relevant Fields	Example Applications	Sample Items	Classic References
Traits	Masculinity and femininity	Social psychology; personality psychology	Gender roles	arrogant, gentle, sociable	Bem (1974)
Risk sources	Dread-inducement and unknowability of potential	Behavioral economics; risk analysis; public policy	Risk behaviors	marijuana, tsunami, hackers	Slovic (1987)
People	Warmth and competence of	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)
Foods	Taste and nutrition	Health psychology; public health policy	Dietary behavior; public health	carrots, tiramisu, celeriac	Raghunathan, Naylor, and Hoyer (2006)
Occupations	Significance and autonomy	Industrial-organizational psychology; labor economics	Career choices; job satisfaction	cab driver, neurosurgeon, historian	Hackman and Oldham (1976)
Brands	Sincerity and excitement of	Marketing; consumer psychology; industrial-organizational psychology	Purchasing behavior; organization-public relations	Home Depot, Comedy Central, ING Direct	Aaker (1997)
Goods	Hedonic and utilitarian value of	Marketing; consumer psychology; psychology of motivation	Purchasing and consumption behavior	chips, vest, hammer	Batra and Ahtola (1990)

RESULTS AND DISCUSSION

The first part of the main analyse the actor applied the test model against the pre-trained model (both models were described above) using the skip-gram technique in order to get the prediction of each judgment domain as vectors. It was then applied ridge regression with regularization hyperparameter (λ set to 10) on the vectors for all but one judgment target along with Pearson correlation coefficients, in which **f+-or** each judgment dimension, was applied leave-one-out cross-validation (LOOCV). As can be seen in Figure 9, this 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 ((Masculine = 0.73 + Feminine = 0.81 + Dread-inducing = 0.88 + Unknowable = 0.86 + Signicance = 0.82 + Autonomy = 0.82 + Warm = 0.79 + Competent = 0.75 + Sincere = 0.59 + Exciting = 0.62 + Tasty = 0.66 + Nutritious = 0.83 + Hedonic = 0.84 and Utilitarian = 0.78) / 14 judgment dimensions = 0.77) , and all fourteen judgments yielding statistically significant positive correlations (all $p < 10–20$).

Thus, it was evaluated the tested model's predictive accuracy with a more straightforward, baseline approach that depends on the relative similarity of a judgment target, using leave-one-out cross-validation. The baseline approach involves learning a linear transformation from the measure of relative vector similarity to the response scale in consideration. It outcomes an average correlation of .30, which is lower than obtained by the vector mapping method. Also, the similarity method outcomes significant correlations only for eleven out of the fourteen tests.

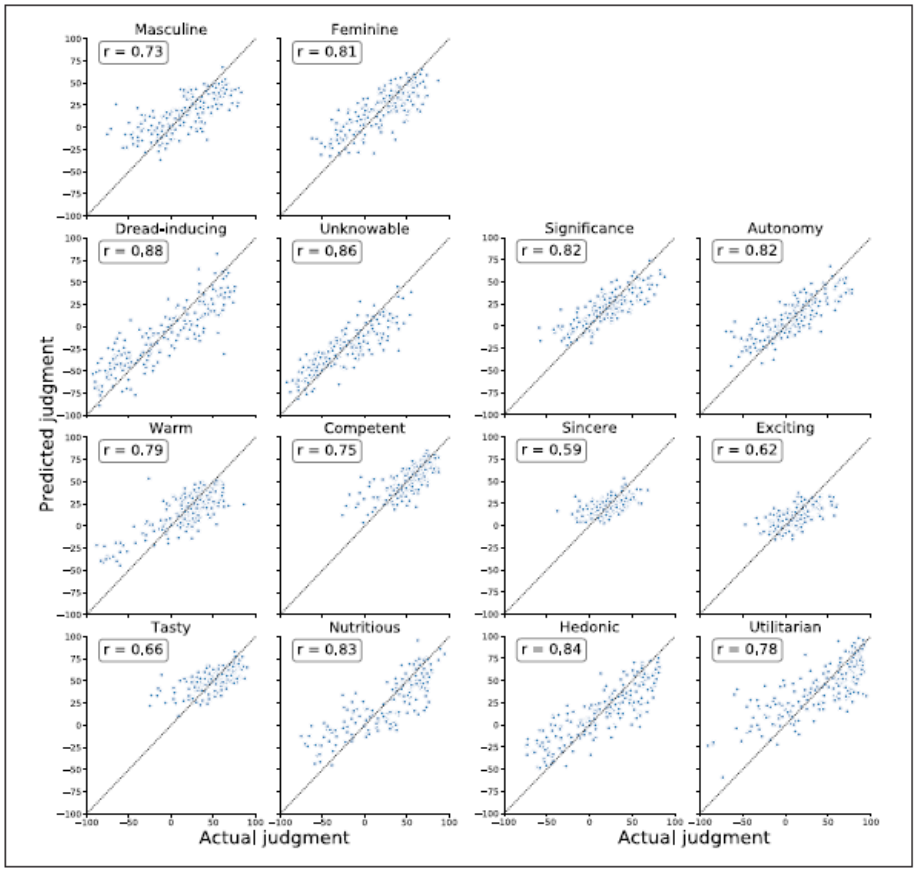

Figure 9 - Scatterplots of actual judgments and predicted judgments using leave-one-out cross-validation for each judgment dimension (Richie et al., 2019).

Both methods' predictive accuracy was compared with human inter-rater reliability, which is largely used to assess word embeddings' ability to model semantic judgments. While human inter-rater reliability came out to 0.60, the

proposed method outcomes an average correlation of 0.77 across judgments, which means that the proposed method predicts human judgments better than individual human judgments.

Moreover, every judgment dimension was split into two sets, calculated the averaged judgment ratings within each set, and computed the correlation between its averages. This process was repeated 100 times. The split-half reliability average outcomes .88 for all judgments dimensions. Figure 10 shows the results of the similarity and mapping method predicted against the leave-one-out cross-validation, inter-subject correlations, and split-half reliabilities for every judgment dimension.

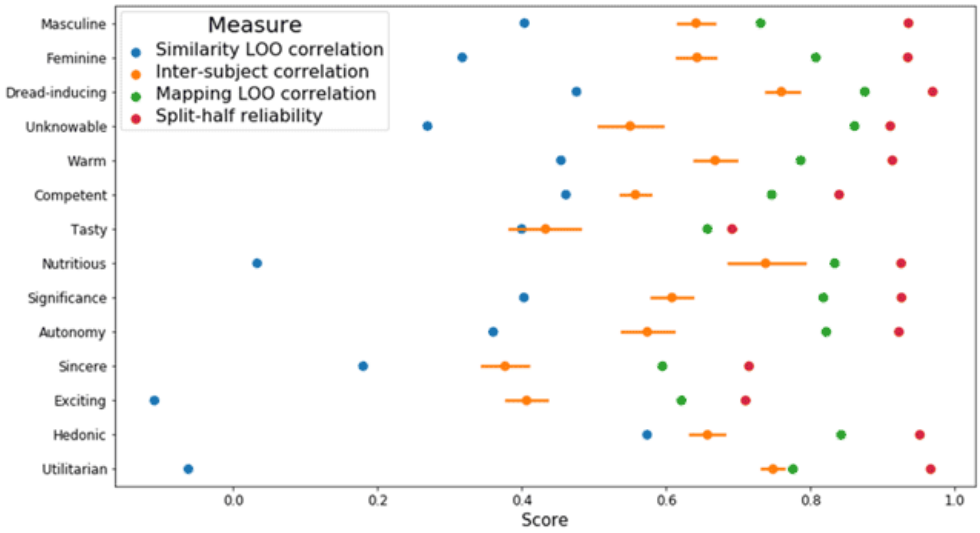

Figure 10 - Results of the similarity and mapping method predicted against the leave-one-out cross-validation, inter-subject correlations, and split-half reliabilities for every judgment dimension (Richie et al., 2019).

Figure 20 allows a comparison of the accuracy of all judgment dimensions, which shows a good variability in performance across the judgment dimensions. While risk source unknowability was predicted with a correlation of .86, food item taste was predicted with a correlation of .66. The variability is likely due to a variety of factors. For instance, judgments may be differently understood by word embeddings.

The analysis above described how the method proposed by Richie et al., 2019, was applied in a pre-trained word2vec embeddings model. As can be seen, this method's predictive accuracy is high, which in some cases, such as Dread-induction getting at 0.88. However, to show this approach's generality, some other pre-trained embedding models, varying both in training algorithms and in the training corpus. The GloVe model was trained on the Common Crawl 840B corpus. FastText on the Common Crawl 143 600B corpus. Paragram-SL999 trained on 1.8B tokens of English Wikipedia, using a large database of paraphrase pairs. GloVe-Common-Crawl embeddings are further trained in a similar manner and on similar external constraints on word-word relations. Moreover, it was used the AllenNLP Python package to obtain ELMo vectors 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". That extracts a vector for the sense of the word relevant to the current domain and judgment dimension (i.e., "bass" as in the fish and not the musical instrument). The second approach computes a vector without any such contextualizing and is thus more analogous to fastText, another character-based word embedding technique that does not provide contextualized embeddings. Thus, Bert-as-a-service Python package to similarly derives contextualized and decontextualized BERT embeddings. In contrast, the Test-model is a GloVe model, containing had 197 traits, 198 risk sources, 162 foods, 169 occupations, and 185 consumer goods. Finally, for each embedding type, it was performed the cross-validation proposed on this models.

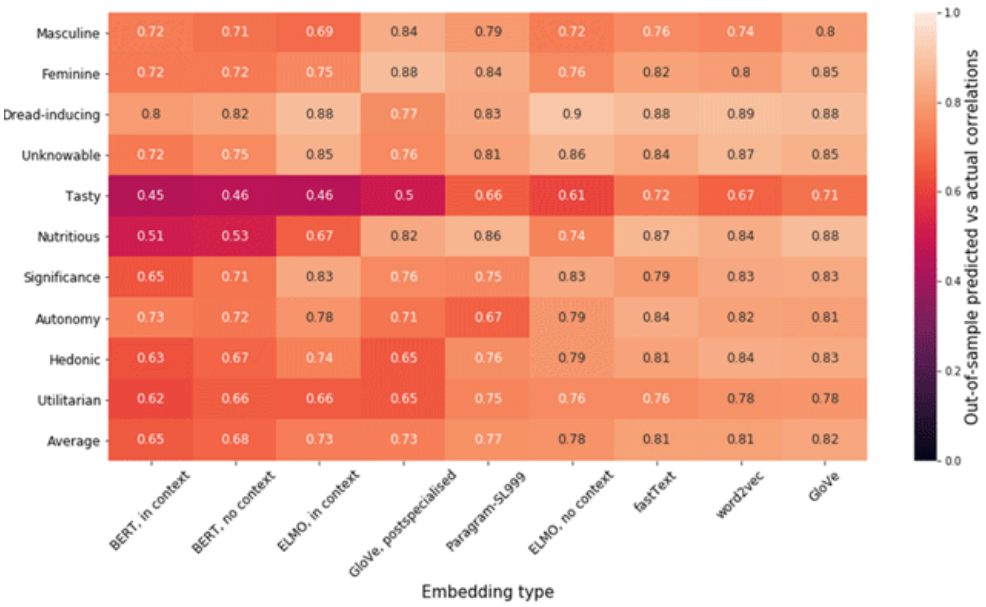

Figure 10 - Pearson correlations between out-of-sample predicted and actual judgments for every judgment dimension except warmth and competence, for different word embedding models (Richie et al., 2019).

As can be seen, the proposed method performance is quite well for many embedding models and judgment dimensions. However, there is such a small variation with embedding types. For instance, masculinity and femininity judgments seem to benefit from embeddings from Paragram-SL999, word2vec, and Glove-Postspec and Glove. However, these algorithms performed worse on the rest of the judgment dimensions, leading to their more insufficient overall accuracy (Richie et al., 2019).

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

Bullinaria, J. A., & Levy, J. P. (2006). Extracting semantic representations from word co-occurrence statistics: A computational study. <https://link.springer.com/content/pdf/10.3758/BF03193020.pdf>.

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

Echizen'ya, H., Araki, K., & Hovy, E. (2019). Word Embedding-Based Automatic MT Evaluation Metric using Word Position Information. In Paper presented at the NAACL-HLT 2019, Minneapolis, Minnesota. June 2 - June 7, 2019. <https://www.aclweb.org/anthology/N19-1186.pdf>.

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems With Evaluation of Word Embeddings Using Word Similarity Tasks.< https://arxiv.org/abs/1605.02276>

Gomez, R. L., Gibert, J. , & Dimosthenis, K. (2019). Chapter 9 - Self-Supervised Learning from Web Data for Multimodal Retrieval, Multimodal Scene Understanding Algorithms, Applications and Deep Learning 2019, Pages 279-306. DOI: https://www.sciencedirect.com/science/article/pii/B9780128173589000159?via%3Dihub.

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.

Hill F., Reichart R. & Korhonen A. (2014). Multi-Modal Models for Concrete and Abstract Concept Meaning.<https://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00183 >

Hollis, G., Lefsrud, L., & Westbury, C, F. (2016). Extrapolating Human Judgments from Skip-gram Vector Representations of Word Meaning. Quarterly journal of experimental psychology (2006). <http://dx.doi.org/10.1080/17470218.2016.1195417>.

Jay, A. (2018). Visualizing machine learning one concept at a time. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. < http://jalammar.github.io/illustrated-word2vec/>.

Lange R.T. (2011) Inter-rater Reliability. In: Kreutzer J.S., DeLuca J., Caplan B. (eds) Encyclopedia of Clinical Neuropsychology. Springer, New York, NY. https://doi.org/10.1007/978-0-387-79948-3_1203.

Li, B., & Han, L. (2013). Distance Weighted Cosine Similarity Measure for Text Classification. In: International Conference on Intelligent Data Engineering and Automated Learning, pp. 611–618. Springer, Berlin (2013).< https://link.springer.com/chapter/10.1007/978-3-642-41278-3_74>.

Ning Liu, Benyu Zhang, Jun Yan, Qiang Yang, Shuicheng Yan, Zheng Chen, Fengshan Bai, and Wei-Ying Ma. 2004. Learning similarity measures in non-orthogonal space. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04). Association for Computing Machinery, New York, NY, USA, 334–341. DOI:https://doi.org/10.1145/1031171.1031240

Karen, S. J. (1994). Natural language processing: a historical review. Ed Zampolli, Calzolari. http://www.mt-<archive.info/Zampolli-1994-Sparck-Jones.pdf>.

Kovacs, B., & Kleinbaum, A. M. (2020). Language-Style Similarity and Social Networks. Psychological science, 31(2), 202–213. https://doi.org/10.1177/0956797619894557.

Jang B., Kim I., & Kim J. W. (2019) Word2vec convolutional neural networks for classification of news articles and tweets. PLoS ONE 14(8): e0220976. https://doi.org/10.1371/journal.pone.0220976.

Jahr, K., Schlich, R., Dragos, K., & Smarsly, K. (2015). Decentralized autonomous fault detection in wireless structural health monitoring systems using structural response data. https://www.researchgate.net/publication/289479445_Decentralized_autonomous_fault_detection_in_wireless_structural_health_monitoring_systems_using_structural_response_data

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.< https://link.springer.com/content/pdf/10.3758/BF03192726.pdf>.

Mikolov, T., Chen, K., Corrado, & G., Dean, J. (2013). Efficient Estimation of Word Representations in VectorSpace. <https://arxiv.org/pdf/1301.3781.pdf>.

Naili, M., Chaibi, A. H., & Ghezala, H. H. B. (2019). Comparative study of word embedding methods in topic segmentation. In Paper presented at the International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France. <https://www.sciencedirect.com/science/article/pii/S1877050917313480/pdf?md5=beea950b0f48b40889edb4491b55c2ef&pid=1-s2.0-S1877050917313480-main.pdf>.

Nagata, H. (1987). The Relativity of Linguistic Intuition: The Effect of Repetition on Grammaticality Judgments. In Paper presented at the Journal of Psychohnguistic Research, Vol. i7, No. 1, 1988. <https://link.springer.com/content/pdf/10.1007/BF01067178.pdf>.

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

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

Richie, R. (2019). Supplementary Information for Predicting high-level human judgment across diverse behavioral domains.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Arzteblatt international*, *107*(44), 776–782. https://doi.org/10.3238/arztebl.2010.0776.

Tzu-Tsung, W. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Journal of Personality and Social Psychology, 48(9), 2839-2846. DOI: https://doi.org/10.1016/j.patcog.2015.03.009.

Waldmann, P. (2019). On the Use of the Pearson Correlation Coefficient for Model Evaluation in Genome-Wide Prediction. Frontiers in Genetics, *899*(10), 1664-8021. DOI: https://www.frontiersin.org/article/10.3389/fgene.2019.00899.

https://towardsdatascience.com/beyond-word-embeddings-part-2-word-vectors-nlp-modeling-from-bow-to-bert-4ebd4711d0ec.

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

Senel, L. K., Utlu, I., Yucesoy, V., & Koc, A., Cukur, T. (2018). Semantic Structure and Interpretability of Word Embeddings. <https://arxiv.org/pdf/1711.00331.pdf >.

Slovic, P. (1987). Perception of risk. Science, 236(4799), 280–285. DOI: https://doi.org/10.1126/ science.356350.