

You Don't Sound Like Me: Analyzing Hiring Based on Pre-hire Language Use

Stanford CS224N Custom Project – **Please Return Graded Feedback**

Samaksh Goyal

Department of Computer Science
Stanford University
sagoyal@stanford.edu

Abstract

This project attempts to model how a job applicant's cultural similarity to other hired applicants is predictive of getting hired into an organization. We use language as a proxy for culture, and apply various natural language processing tools to evaluate applicant response similarity. The applicants responses were from a mid-sized technology company that agreed to embed text based questions at the pre-hire stage, which allowed us to understand how culture specifically influences the hiring process independent of observable candidate traits. We found that classic TF-IDF, out performed Word2Vec, Universal Sentence Encoder and Sentence BERT language models for estimating a statistically significant measure of cultural similarity. Together these results give us new insights into how culture can be empirically modeled for organizations.

1 Key Information to include

- Mentors:
 - Amir Goldberg - Associate Professor of Organizational Behavior Stanford Graduate School of Business (amirgo@stanford.edu)
 - Sameer B. Srivastava - Associate Professor and Harold Furst Chair in Management Philosophy and Values Haas School of Business (sameersriv@berkeley.edu)
- External Collaborator: Sarah K. Stein - Data Scientist Paradigm Strategy Inc., Ph.D. in Organizational Behavior Stanford Graduate School of Business (skstein2@gmail.com)

2 Introduction

Culture actively shapes an individual's behavior and defines social structure within communities. In sociology it is an exciting topic of interest as understanding, evaluating and manipulating culture can lead to novel group insights [1]. Taken a step further, the new trend of social engineering has recently amassed generous attention as (corporate) organizations try to analyze their employees' cultural fit [2, 3].

Yet arguably the most important cultural match an organization must make with individuals is during the hiring process. Indeed, organizations are almost taking a gamble on the individual in question and make an evaluative decision to employ the individual based solely on their displayed characteristics. This gives candidates the upper hand as they are able to use self-presentation tactics to make themselves appear more attractive and only reveal partial information [4]. This asymmetry creates two schools of thought for corporations. The first is to hire based on traits of the applicant that are not relevant to the job and are associated with the applicants underlying personality (commonly thought of as the "would I be willing to spend 7 hours stuck at an airport with this person" test). The second is to hire individuals with a sub-par skill-set (in the hope that they can develop the necessary ones), but those who are able to connect with cultural work environment.

However accurately measuring this "cultural fit" based on pre-hire characteristics can be biased. For example hiring managers' are often swayed by easily observable cultural traits like hairstyle and attire which are noisy signals at best. Furthermore many managers' select for shared sociodemographic traits like hobbies and lifestyle preferences which can lead them to overlook culturally dissimilar but otherwise promising candidates [5].

Instead, this paper relies on the assumption that language is the raw manifestation of culture, a growing body of work led by Amir Goldberg and Sameer Srivastava [6, 7, 8, 9]. To do so we placed short essay questions into the hiring process of a mid-sized technology firm and developed several NLP models of cultural similarity between job applicants and previously hired employees. (Note these responses were not made available to the candidate selection committee nor interviewers). This method now allows us to identify how language use at the pre-hire stage relates to a job candidate's likelihood of being hired independent of the traits that are observable to candidate selection committee and interviewers.¹

The above methodology, developed by Stein [10], tested the hypothesis using only TF-IDF to model applicant response. For this paper, I extend on her work and test several word embedding models hoping to capture more cultural nuance in applicant responses. To do so I develop an unsupervised method to assess cultural similarity based on responses to essay questions.

3 Related Work

The section summarize the various semantic textual similarity methods used to measure linguistic similarity. In Stein's prior work she evaluated TF-IDF. However with the advent of finer language models, we expanded to word embeddings. Yet, this led us to the challenge of constructing response (sentence) level embedding with only word vectors. As a naïve approach we constructed a sentence embedding as an average of all word embeddings in the sentence (where word embeddings were optimized for being averaged). An alternative approach was Word Mover Distance that allowed us to compare sentences without formally constructing a sentence vector. This method summed across the distance each word in one sentence needed to move to transform to the other sentence. Finally we also evaluated two sentence level embeddings Universal Sentence Encoder and Sentence BERT, which had sentence embeddings tailored for transfer tasks such as semantic textual similarity,

3.1 TF-IDF

For lengthy documents it is often important to reduce the dimensionality of the text into components, which can then be compared for similarities across individuals. In the case of our pre-hire essay questions, their relatively short length (a few sentences on average) and their narrow focus (responding to a direct inquiry) made dimensionality reduction less relevant. As such, we chose to analyze the responses to these questions using similarities in individual unigram frequencies under the assumption that word choice itself would be informative about culture.²

3.2 Average Word Embedding

After finding word embeddings, it is difficult to clearly see how the idea can be abstracted to find embeddings for larger phrases such as those at the sentence level. One idea is to simply average over each word in the sentence and have it be representative of a sentence vector [11]. Kenter et al. follow the same procedure as Mikolov et al. strategy at the sentence level, where they try to predict a sentence from its adjacent sentences. They compute sentence embeddings by averaging word embeddings, but are optimizing word embeddings directly for the purpose of being averaged.

More concretely for a pair of sentences (s_i, s_j), let $p(s_i, s_j)$ be a probability that reflects how likely it is for the sentences to be adjacent to another in the training data. Then compute the probability

¹I want be clear that methodology to model a candidate's linguistic response as proxy for culture was established by my mentors Goldberg and Srivastava and is not my own original idea. Furthermore the experimental design of placing essay questions that were unobserved by the selection committee was thought of by Sarah K. Stein. All of the sociological background and relevant details surrounding the biases of hiring in organizations was also founded by Sarah K. Stein.

²From Sarah K. Stein's prior work on TF-IDF model selection

$p(s_i, s_j)$ using a softmax function. $p_\theta(s_i, s_j) = \frac{e^{\cos(s_i^\theta, s_j^\theta)}}{\sum_{s_l \in S} e^{\cos(s_i^\theta, s_l^\theta)}}$. Note that again while we would like to iterate over all sentences in the denominator, once again we use negative sampling to select a subset of S^+ . The loss function is categorical cross-entropy:

$$L = - \sum_{s_j \in \{S^+ \cup S^-\}} p(s_i, s_j) \cdot \log(p_\theta(s_i, s_j))$$

where $p(\cdot)$ is the target probability the network should produce and $p_\theta(\cdot)$ is the prediction it estimates based on parameter θ . The target distribution is uniform with:

$$p(s_i, s_j) = \begin{cases} \frac{1}{|S^+|}, & \text{if } s_j \in S^+ \\ 0, & \text{if } s_j \in S^- \end{cases}$$

3.3 Word Mover Distance

Another method purposed was to find the sentence similarity(distance) between text documents is to first take traditional word embeddings, and then find the minimum cumulative distance that words from document A need to travel to match exactly the position of document B [12].

More concretely, the “travel cost” between two text documents is defined as the following. By starting with $\mathbf{d}$ and $\mathbf{d}'$ being the nBOW document representations, let each word i in $\mathbf{d}$ to be transformed into any word in $\mathbf{d}$ (in total or in parts). Then let $T \in \mathbb{R}^{n \times n}$ be a (sparse) flow matrix where $T_{ij} \geq 0$ denotes how much of word i in $\mathbf{d}$ travels to word j in $\mathbf{d}'$. Then the total cost is the cost required to move all words from $\mathbf{d}$ to $\mathbf{d}'$: $\sum_{ij} T_{ij} c(i, j)$.

3.4 Universal Sentence Encoder

Google researchers Cer et al. created general purpose "universal" sentence embedding that could be used several for transfer applications (one of which was Semantic Textual Similarity) [13].

The embeddings were created through two methods. The first was using making sentence embeddings using the encoding sub-graph of the transformer architecture [14]. Using attention, context aware representations of the words took into account the ordering as well. Then fixed length encoding vectors were made by computing the element-wise sum of each word in the sentence and normalizing with by square root of the length of the sentence.

The second encoding model was deep averaging networks where the word and bi-gram embeddings were first averaged and then forward propagated through deep neural network to produce sentence embeddings [15].

The training was done for multi-task learning where single encoding models were used for multiple downstream tasks.

3.5 Sentence BERT

While BERT had hit new performance thresholds on sentence classification and sentence-pair regression tasks, Reimers et al. recognized that the cross-encoder BERT used (have two sentences pass to the transformer network and then predicting a target value) was not scalable in terms of compute time [16].

The paper uses a pre-trained BERT embedding to fine-tune sentence embeddings. First pooling operation to the output of BERT is applied (either the CLS token, computing the mean of all output vectors, or computing a max-over-time of the output vectors). Then to tune BERT a siamese and triplet networks updates the weights so that the output sentence embeddings are semantically meaningful and had close cosine-similarity [17].

4 Approach

4.1 Linguistic Similarity

To measure Linguistic (Semantic) Similarity between applicants, we tried several methods of modeling response similarities³. Below is the general procedure:

1. Construct pairwise similarities between all applicant responses (based on some similarity function sim). Resulting in a matrix as such:

$$S = \begin{bmatrix} 0 & sim(\theta)_{12} & \dots & sim(\theta)_{1n} \\ sim(\theta)_{21} & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ sim(\theta)_{n1} & \dots & \dots & 0 \end{bmatrix}$$

2. Construct an independent pre-hire fit measure, S_H , by multiplying the similarity matrix, S , by a vector of binary hiring outcomes, $H = [0, 1]$. This resulted in a variable measuring the degree to which each individual's language was similar to the language used by the group of hired individuals (excluding himself or herself when appropriate).

$$H = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}, S_H = S * h = \begin{bmatrix} 0 & sim(\theta)_{12} & \dots & sim(\theta)_{1n} \\ sim(\theta)_{21} & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ sim(\theta)_{n1} & \dots & \dots & 0 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}$$

For each method of modeling response similarity we have different procedures:

4.1.1 TF-IDF

For each applicant response a Term Frequency vector was made and then cosine similarities ($sim = cos$) were computed between all applicant responses. This will be our baseline against which to compare all latter embedding models.

4.1.2 Word2Vec Average Word

Either all essay responses or firm emails were used as the training corpus to create CBOW Word2Vec embeddings. Then for each response, we found the average word embedding by iterating over all words in the response to find centroid. Then cosine similarities ($sim = cos$) were computed between all applicant average responses.

4.1.3 Word2Vec Word Mover Distance

Either all essay responses or firm emails were used as the training corpus to create CBOW Word2Vec embeddings. Then for each response we used Word Mover distances between applicant responses to model sentence similarity ($sim = WMD$).

4.1.4 Universal Sentence Encoder

Pre-trained fixed length sentence embedding from Google were downloaded and retrieved for each applicant response. Then cosine similarities ($sim = cos$) were computed between all USE sentence encoding of applicant responses.

³I would like to mention that this part of the paper is my original contribution to the project. I researched, found and developed ways of using existing NLP work based on semantic textual similarity to apply to this project.

4.1.5 Sentence BERT

Pre-trained fixed length sentence embedding from Siamese BERT network were downloaded and retrieved for each applicant response. Then cosine similarities ($sim = cos$) were computed between all SBERT sentence encoding of applicant responses.

4.2 Generalized Logistic Regression

We ran a generalized logistic regression with the dependent variable of hiring outcome based on S_H and other controls X .

$$\ln \left(\frac{P(H = 1)}{1 - P(H = 1)} \right) = B_0 + B_1 S_H + B_2 X$$

5 Experiments

5.1 Data

The data comes from a mid-sized technology company ⁴ that placed optional essay questions into its job application process. The company received 79,192 job applications, 17% of which chose to respond to at least one question. This gave us a total of 11,587 individuals, 2,128 of whom were interviewed and 353 of whom were hired. For the purposes of this report we only focus on following question answered by applicant about work preferences fit. *Note the for making word embeddings we stemmed and removed stop words.

Sample Question and Answer	
Question	In the space below, describe your ideal work environment.
Answer	My ideal work environment is collaborative and open, where I can both have a voice in what my team produces, but more importantly learn from my coworkers and consistently evolve.
Processed Answer	ideal work environ collabor open can voic team produc import learn cowork consist evolv

The data also includes applicants specific data such as if they were applying for a managerial role (indicator), the log word count of their response, gender, age, if they were sourced through recruiter or employee referral (indicator). These were used as controls.

In addition we also had a corpus of all emails sent within the corporation. This corpus was used as the embedding space for some experiments.

5.2 Evaluation method

To evaluate if the proposed measure for linguistic similarity score was significant factor for regressing on hiring outcome we look at the p-values calculated for each regression on similarity score parameter.⁵

Furthermore to ensure that the the similarity measure is capturing the likelihood of hiring, we built multiple models to tell us if there were omitted variables in our regression (as seen by the addition of covariates in our regression). The first regression (1) is looking at the language difference between hired and non-hired controlling for their desired role. The second regression (2) add demographic controls like age and gender. The third regression (3) add human capital controls such as applicant's

⁴Company name concealed due to NDA.

⁵Unlike traditional Machine Learning projects which have a train/dev/test sets for data, this methodology is not appropriate for this project. We are only testing in-group samples of all applicants when constructing cultural similarity score and so the only metric of concern is the p-value and the associated coefficient. Unlike traditional NLP project we do not care for the "accuracy" so to speak of the prediction on whether someone is hired or not, rather we care more about the numerical meaning and significance of the cultural similarity parameter.

from high-status organizations and top-tier universities. The fourth regression (4) also adds employee referral.

5.3 Experimental details

Word2vec embeddings were created using the response corpus, and the email corpus using the Gensim library. The model parameters were: window=6, size=100, alpha=0.03, negative=20. The Gensim package Phrases and Phraser were also used so that common phrases like "new" and "york" would be treated as the collective "new_york."

Since the word mover distance need to be calculated between all $\binom{2128}{2}$ response pairs the python multiprocessing library was used to multi-thread the operation.

For Sentence BERT, bert-large-nli-stsb-mean-tokens were used to seek sentence embeddings as they produced the highest performance semantic text similarity benchmark of 79.19.

5.4 Results

Below are the results for regression (1) - no controls and regression (4) - full controls for each similarity model. These are the coefficients on the cultural (linguistic) similarity parameter and the p-value, for each of our language models. The full results for all regression is available in appendix (Tables 2-9).

Table 1: Cultural (Linguistic) Similarity

	(1)	(4)
TF-IDF	1.217**	1.209*
Word2Vec (Response Corpus) Average Word	0.373*	0.540
Word2Vec (Response Corpus) WMD	0.148*	0.283
Word2Vec (Email Corpus) Average Word	0.858*	0.916
Word2Vec (Email Corpus) WMD	0.652*	0.752
Universal Sentence Encoder	4.264	2.117
Sentence BERT	3.443	2.531
<i>Associated p-values:</i>		
	*p<0.05; **p<0.01	

Contrary what was expected, TF-IDF actually produced the most statistically significant results for the cultural similarity measure. This result is held for all regressions (again see Tables 2-9). Even ordinary Word2Vec embeddings prove to be more significant indicators of cultural similarity, then the more novel transformer and siamese networks made with USE and SBERT.

Furthermore even after all controls are added in regression (4) all measures lose significance except for TF-IDF.

6 Analysis

Taking a step back, and recalling the assumption that language was being used as proxy for culture we can assess the above results holistically.

TF-IDF that most coarse measure (simply checking for word co-occurrences) produces the most significant measure for cultural linguistic similarity. A possible reasoning for this could be since the essay question was already so narrow it filtered for a specific type of candidate and exact word matches were more likely to be predictive of a candidate's hiring outcome. Then the use of more abstract representations like word embeddings, Universal Sentence Encoder and Sentence BERT would not prove to be any more helpful.

Along this line, the vocabulary of responses from all candidates is very limited compared to the global vocabulary of all words present in the embedding. This could mean that the over-exaggerated differences seen between term frequencies is more representative of a candidate's cultural similarity than more abstract word vectors.

Furthermore by using word embeddings we must take a two-step jump before we can start interpreting applicant response interactions. First embedding representations of words need to be made and then an artificial sentence representation needs to be constructed (normalizing over all words in the sentence). However TF-IDF does not have that hindrance, as it avoids sentence level re-aggregation as it is already present in the term frequencies. It is possible just the frequencies of certain words alone is more representative of hiring. The hypothesis being that if a candidate states a few key "trigger" words they are immediately more likely to be hired even if another candidate says a similar phrase with the same meaning but different words. This follows from Rivera [5] showing qualitative evidence that cultural matching is often biased for lifestyle characteristics and links. In the context of sociology and organizational behavior this is quite interesting as it means that selection committees could have bias for candidates with a particular vocabulary. However this hypothesis would need to be fully examined.

Another possible reasoning for the above results could be in our approach of constructing S_H (the degree to which each individual's language was similar to the language used by the group of hired individuals). We used the method of multiplying the similarity matrix S with the hiring outcomes $H = [0, 1]$ because it gave us a relationship between a candidate and all hired applicants. However it is possible a different model could be used to represent group to individual relationships.[18].

An interesting result that can be seen in Tables 2-9 is that Manager (an indicator if a candidate as applying for a management position) remains a significant feature across all models and regressions. This makes sense since our measure is never detracting or adding bias to other non-linguist controls.

7 Conclusion

In this project we analyze if cultural similarity to previously hired job applicants is predictive of a job candidate's hiring outcome, even after controlling for the applicant's human and social capital. To do this we construct a proxy measure for cultural similarity, using language, and rigorously test if the measure is significant between a job applicant and all hired applicants.

We found that the best language measure for cultural similarity is TF-IDF, which beat out all word and sentence embedding based models. Even though this measure might be perceived as being too coarse, this characteristic in fact might be reason for its significance. It is able to capture the restricted candidate response vocabulary to be a significant measure of cultural similarity.

Some major achievements of this project is that we were able to have access to a unique dataset of applicant responses that were **not** used by the selection committee for hiring criteria. This allowed to independently test for cultural work preference similarities between all applicants regardless of observable traits. Furthermore no such prior work has been done to empirically evaluate cultural significance during the hiring process, all other work was either qualitatively or had the prior mentioned biases.

A limitation of this work is that the essay response answers might be too narrow to accurately model all of cultural aspects of an candidate's work preferences. Thus next we could broaden the scope of plausible answers.

Some broader implication of this work imply that hiring committees need to be cognizant of the culture they are selecting for when they find candidates. This is important as now many corporations, specifically those in tech and finance, are pushing to diversify their organizations. An application of these results could mean that firm seeking to hire for cultural fit might place comparable free response questions to the ones we used in the job application process and then use a similar method to identify like minded individuals.

References

- [1] Theresa Schmiedel, Oliver Müller, and Jan vom Brocke. Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. *Organizational Research Methods*, 22(4):941–968, 2019.
- [2] Waldo Rocha Flores and Mathias Ekstedt. Shaping intention to resist social engineering through transformational leadership, information security culture and awareness. *computers & security*, 59:26–44, 2016.
- [3] Rameshwar Dubey, Angappa Gunasekaran, Stephen J Childe, Thanos Papadopoulos, Benjamin Hazen, Mihalios Giannakis, and David Roubaud. Examining the effect of external pressures and organizational culture on shaping performance measurement systems (pms) for sustainability benchmarking: Some empirical findings. *International Journal of Production Economics*, 193:63–76, 2017.
- [4] Murray R Barrick, Jonathan A Shaffer, and Sandra W DeGrassi. What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, 94(6):1394, 2009.
- [5] Lauren A Rivera. Hiring as cultural matching: The case of elite professional service firms. *American sociological review*, 77(6):999–1022, 2012.
- [6] Matthew Corritore, Amir Goldberg, and Sameer B Srivastava. Duality in diversity: Cultural heterogeneity, language, and firm performance. 2018.
- [7] Amir Goldberg, Sameer B Srivastava, V Govind Manian, William Monroe, and Christopher Potts. Fitting in or standing out? the tradeoffs of structural and cultural embeddedness. *American Sociological Review*, 81(6):1190–1222, 2016.
- [8] Sameer B Srivastava, Amir Goldberg, V Govind Manian, and Christopher Potts. Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science*, 64(3):1348–1364, 2018.
- [9] Gabriel Doyle, Amir Goldberg, Sameer Srivastava, and Michael C Frank. Alignment at work: Using language to distinguish the internalization and self-regulation components of cultural fit in organizations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 603–612, 2017.
- [10] Sarah Kathryn Stein. *DISTINGUISHING ROUND FROM SQUARE PEGS UNDERSTANDING HIRING BASED ON PRE-HIRE LANGUAGE USE*. PhD thesis, 2018.
- [11] Tom Kenter, Alexey Borisov, and Maarten De Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*, 2016.
- [12] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015.
- [13] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [15] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, 2015.
- [16] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

- [17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [18] Marina Tulin, Thomas V Pollet, and Nale Lehmann-Willenbrock. Perceived group cohesion versus actual social structure: A study using social network analysis of egocentric facebook networks. *Social science research*, 74:161–175, 2018.

A Appendix

A.1 TF-IDF

Table 2:

	<i>Hiring Odds Ratio</i>			
	(1)	(2)	(3)	(4)
Linguistic Similarity	1.217 t = 2.997**	1.245 t = 2.905**	1.261 t = 3.050**	1.209 t = 2.435*
Manager	0.254 t = 6.420***	0.232 t = -5.927***	0.238 t = -5.813***	0.252 t = -5.547***
Log Word Count	1.162 t = 1.594	1.120 t = 1.056	1.115 t = 1.002	1.084 t = 0.728
Gender (Female)		0.814 t = 1.475	0.833 t = 1.306	0.873 t = 0.953
Age		1.022 t = 2.553*	1.021 t = 2.514*	1.018 t = 1.882
High-status Organization			0.835 t = 0.847	0.879 t = 0.600
Top-tier University			0.652 t = 3.152**	0.675 t = 2.675**
Sourced through Recruiter			12.148 t = 5.417***	15.027 t = 5.881***
Employee Referral				8.211 t = 12.449***
Constant	0.024 t = 13.286***	0.016 t = 10.306***	0.019 t = 9.693***	0.018 t = 9.457***
Observations	10,281	8,072	8,072	8,072
Log Likelihood	-1,397.5	-1,084.5	-1,068.8	-1001.6
Akaike Inf. Crit.	2,803.0	2,181.1	2,155.5	2,043.2

Note:

*p<0.05; **p<0.01; ***p<0.001

A.2 Word2Vec (Response Corpus) Average Word

Table 3:

	<i>Hiring Odds Ratio</i>			
	(1)	(2)	(3)	(4)
Linguistic Similarity	0.373 t = -2.390*	0.452 t = -1.711	0.423 t = -1.834	0.540 t = -1.291
Manager	0.258 t = -6.341***	0.234 t = -5.878***	0.239 t = -5.777***	0.257 t = -5.459***
Log Word Count	1.124 t = 1.379	1.152 t = 1.476	1.163 t = 1.575	1.171 t = 1.612
Gender (Female)		0.855 t = -1.121	0.873 t = -0.970	0.913 t = -0.636
Age		1.022 t = 2.545*	1.022 t = 2.534*	1.017 t = 1.853
High-status Organization			0.799 t = -1.039	0.837 t = -0.812
Top-tier University			0.669 t = -2.954**	0.698 t = -2.600**
Sourced through Recruiter			13.195 t = 5.510***	16.606 t = 5.998***
Employee Referral				8.585 t = 12.715***
Constant	0.054 t = -7.587***	0.027 t = -7.546***	0.032 t = -7.095***	0.022 t = -7.525***
Observations	10,281	8,072	8,072	8,072
Log Likelihood	-1,371.265	-1,070.779	-1,055.047	-995.751
Akaike Inf. Crit.	2,750.530	2,153.558	2,128.095	2,011.503

Note:

*p<0.05; **p<0.01; ***p<0.001

A.3 Word2Vec (Response Corpus) WMD

Table 4:

	<i>Hiring Odds Ratio</i>			
	(1)	(2)	(3)	(4)
Linguistic Similarity	0.148 t = -2.246*	0.180 t = -1.782	0.168 t = -1.835	0.283 t = -1.293
Manager	0.257 t = -6.359***	0.234 t = -5.891***	0.239 t = -5.792***	0.257 t = -5.471***
Log Word Count	1.125 t = 1.388	1.153 t = 1.482	1.164 t = 1.580	1.171 t = 1.610
Gender (Female)		0.860 t = -1.081	0.879 t = -0.920	0.918 t = -0.598
Age		1.021 t = 2.531*	1.022 t = 2.510*	1.017 t = 1.841
High-status Organization			0.800 t = -1.034	0.836 t = -0.819
Top-tier University			0.670 t = -2.950**	0.698 t = -2.600**
Sourced through Recruiter			12.932 t = 5.470***	16.365 t = 5.970***
Employee Employee Referral				8.575 t = 12.707***
Constant	0.031 t = -13.100***	0.017 t = -10.973***	0.020 t = -10.438***	0.016 t = -10.463***
Observations	10,281	8,072	8,072	8,072
Log Likelihood	-1,371.518	-1,070.562	-1,054.960	-995.721
Akaike Inf. Crit.	2,751.036	2,153.125	2,127.920	2,011.442

Note:

*p<0.05; **p<0.01; ***p<0.001

A.4 Word2Vec (Email Corpus) Average Word

Table 5:

	<i>Hiring Odds Ratio</i>			
	(1)	(2)	(3)	(4)
Linguistic Similarity	0.858 t = -1.981*	0.877 t = -1.488	0.871 t = -1.554	0.916 t = -0.987
Manager	0.257 t = -6.351***	0.234 t = -5.888***	0.239 t = -5.787***	0.257 t = -5.464***
Log Word Count	1.125 t = 1.391	1.153 t = 1.488	1.164 t = 1.583	1.171 t = 1.612
Gender (Female)		0.857 t = -1.103	0.876 t = -0.946	0.917 t = -0.606
Age		1.021 t = 2.533*	1.022 t = 2.515*	1.017 t = 1.832
High-status Organization			0.801 t = -1.026	0.838 t = -0.806
Top-tier University			0.670 t = -2.946**	0.698 t = -2.598**
Sourced through Recruiter			12.972 t = 5.478***	16.401 t = 5.977***
Employee Referral				8.617 t = 12.735***
Constant	0.048 t = -7.782***	0.025 t = -7.581***	0.030 t = -7.170***	0.020 t = -7.624***
Observations	10,281	8,072	8,072	8,072
Log Likelihood	-1,372.295	-1,071.181	-1,055.584	-996.126
Akaike Inf. Crit.	2,752.590	2,154.362	2,129.168	2,012.252

Note:

*p<0.05; **p<0.01; ***p<0.001

A.5 Word2Vec (Email Corpus) WMD

Table 6:

	<i>Hiring Odds Ratio</i>			
	(1)	(2)	(3)	(4)
Linguistic Similarity	0.652 t = -2.128*	0.673 t = -1.725	0.666 t = -1.761	0.752 t = -1.232
Manager	0.257 t = -6.360***	0.234 t = -5.892***	0.239 t = -5.793***	0.257 t = -5.472***
Log Word Count	1.125 t = 1.394	1.153 t = 1.488	1.164 t = 1.584	1.171 t = 1.611
Gender (Female)		0.861 t = -1.074	0.880 t = -0.911	0.919 t = -0.590
Age		1.021 t = 2.524*	1.021 t = 2.502*	1.017 t = 1.835
High-status Organization			0.800 t = -1.029	0.836 t = -0.817
Top-tier University			0.670 t = -2.947**	0.698 t = -2.599**
Sourced through Recruiter			12.878 t = 5.462***	16.318 t = 5.964***
Employee Referral				8.588 t = 12.717***
Constant	0.031 t = -13.139***	0.017 t = -10.987***	0.020 t = -10.454***	0.016 t = -10.473***
Observations	10,281	8,072	8,072	8,072
Log Likelihood	-1,371.841	-1,070.689	-1,055.126	-995.810
Akaike Inf. Crit.	2,751.682	2,153.377	2,128.251	2,011.621

Note:

*p<0.05; **p<0.01; ***p<0.001

A.6 Universal Sentence Encoder

Table 7:

	<i>Hiring Odds Ratio</i>			
	(1)	(2)	(3)	(4)
Linguistic Similarity	4.264 t = 1.947	3.227 t = 1.378	3.316 t = 1.399	2.117 t = 0.861
Manager	0.257 t = -6.364***	0.233 t = -5.896***	0.239 t = -5.782***	0.253 t = -5.524***
Log Word Count	1.149 t = 1.409	1.161 t = 1.337	1.162 t = 1.337	1.140 t = 1.138
Gender (Female)		0.833 t = -1.306	0.855 t = -1.117	0.897 t = -0.764
Age		1.022 t = 2.578**	1.022 t = 2.555*	1.018 t = 1.878
High-status Organization			0.822 t = -0.921	0.867 t = -0.663
Top-tier University			0.655 t = -3.128**	0.679 t = -2.819**
Sourced through Recruiter			11.372 t = 5.283***	14.280 t = 5.780***
Employee Referral				8.386 t = 12.587***
Constant	0.014 t = -15.547***	0.009 t = -11.896***	0.011 t = -11.307***	0.011 t = -10.753***
Observations	10,722	8,385	8,385	8,385
Log Likelihood	-1,400.585	-1,087.971	-1,072.617	-1,014.380
Akaike Inf. Crit.	2,809.171	2,187.942	2,163.234	2,048.759

Note:

*p<0.05; **p<0.01; ***p<0.001

A.7 Sentence BERT

Table 8:

	<i>Hiring Odds Ratio</i>			
	(1)	(2)	(3)	(4)
Linguistic Similarity	3.443 t = 1.714	3.616 t = 1.548	4.019 t = 1.656	2.531 t = 1.094
Manager	0.251 t = -6.469***	0.229 t = -5.970***	0.235 t = -5.859***	0.251 t = -5.571***
Log Word Count	1.196 t = 1.989*	1.179 t = 1.619	1.175 t = 1.567	1.142 t = 1.264
Gender (Female)		0.834 t = -1.305	0.854 t = -1.124	0.896 t = -0.771
Age		1.021 t = 2.539*	1.022 t = 2.517*	1.018 t = 1.858
High-status Organization			0.825 t = -0.906	0.869 t = -0.652
Top-tier University			0.653 t = -3.149**	0.677 t = -2.837**
Sourced through Recruiter			11.584 t = 5.318***	14.454 t = 5.803***
Employee Referral				8.357 t = 12.561***
Constant	0.011 t = -12.875***	0.007 t = -10.509***	0.008 t = -10.095***	0.009 t = -9.550***
Observations	10,722	8,385	8,385	8,385
Log Likelihood	-1,400.998	-1,087.693	-1,072.186	-1,014.140
Akaike Inf. Crit.	2,809.995	2,187.385	2,162.371	2,048.279

Note:

*p<0.05; **p<0.01; ***p<0.001