SUBMISSION TYPE

Symposium

TITLE

Data-Lite Discovery: Computational/ML Methods to Make the Most Use of Available Data

SHORTENED TITLE

Data-Lite Discovery

ABSTRACT

We showcase a variety of ways that researchers can use pre-learned information from machine learning models and synthetic data to generate novel discoveries and supplement existing datasets. These methods include training models to detect inattention without any human inattention data, approximating humans’ mental perceptions using natural language inferences models, consolidating overlapping constructs through examining item embeddings, and training more robust classification models by augmenting text. These approaches show researchers and practitioners how to be able to do more with less.

CITATIONS

Hernandez, I. (Chair) & Song, Q.C. (Discussant). (2023). Data-Lite Discovery: Computational/ML Methods to Make the Most Use of Available Data [Symposium]. Society for Industrial and Organizational Psychology Annual Conference, Boston, MA, United States.

Nie, W. (2023). Detecting Careless Responding with Machine Learning Models

Trained by Semi-Synthetic Data

Delice, F, Hernandez, I., & Foti, R. (2023). The AI Mirror: Inferring Leader Perceptions Across Demographics using Natural Language Processing

Acton. B., Li, M. & Oh, C. (2023). Leadership Construct Consolidation via NLP: Identifying the Overlap between Conceptually Related Items

Meyer, J., & Koenig, N. (2023). Experimenting with the Dark Arts of Data Augmentation for Natural Language Processing

WORD COUNT

4694

Data-Lite Discoveries: Computational and Machine Learning Approaches to Minimize the Need for Additional Data Collection

Ivan Hernandez

Virginia Tech

I-O psychologists have highlighted the opportunities for machine learning to transform the way assessments are scored (Speer et al., 2022), scales are developed (Hernandez & Nie, 2022), and interviews are evaluated (Thompson et al., 2021). These methods often require hundreds or thousands of annotated examples to implement and derive insights. However, collecting additional data can be time-consuming and costly for researchers and practitioners alike. Therefore, there is a pressing need for research that uses samples efficiently and provides ways to address potential data acquisition challenges, by making use of the information available.

This symposium’s primary intended contribution is to hopefully reshape how I-O researchers and practitioners think of what data are available to derive insights. The papers in this symposium offer a different perspective of machine learning of broad interest that differs from the large collected training data-centric approaches historically discussed in machine learning conversations. Rather, these approaches highlight how many pre-trained deep neural network models have generalizable associations between concepts already consolidated within their internal architecture, which can be extracted to learn about people and associations between items. Additionally, researchers with some available data can expand on that data by synthesizing novel cases that contain informative variations. These synthetic examples can improve the robustness of machine learning models compared to approaches that relied only on the original collected data. Collectively, these projects highlight ways that researchers and practitioners can maximize the data they already have. These approaches can also be beneficial to students and early career researchers, who may not have the access to large datasets needed for training machine learning models, and desire more accessible ways to generate new discoveries through machine learning.

The symposia bring together a diverse collection of researchers from industry and academia who illustrate ways to derive insights by going beyond the original data collected, and obtaining insights in unique ways via machine learning.

In the first paper, **Weiwen Nie** highlights how computers can replicate psychological processes of interest, such as inattentive responding. By using computational approaches to simulate responses that are inattentive, and then using that simulated data in tandem with archival “attentive” data (which is easier to obtain from an organization), a machine learning model was able to outperform other commonly used metrics at distinguishing attentive from inattentive individuals.

The second paper by **Fabrice Delice** and colleagues illustrates how large-language models contain the psychological perceptions of trait associations within their embeddings, and how these perceptions align with reported perceptions across a variety of contexts. They then leverage these natural language inference models to examine how societal perceptions of leadership/followership vary between privileged and marginalized group statuses.

The third paper by **Bryan Acton** and colleagues harnesses the information contained within existing scale items and pre-trained sentence embedding models to resolve construct contamination issues within leadership research. Traditionally, construct contamination is addressed through factor analytic approaches, which would require respondents to complete all possible leadership items to construct the necessary covariance matrix between items. Their approach leverages the ability of sentence embedding models to quantify the semantic similarity between items and discover the perceptual equivalence of the underlying meaning between items.

The fourth paper by **Joe Meyer** and Nick Koenig demonstrates that researchers interested in training text classification models have a variety of tools at their disposal to further maximize the insights available from their existing data. This presentation contributes to the broad community by showcasing the idea of “data augmentation methods” from computer science. These methods, which are not widely used in I-O psychology, include synonym replacement, contextualized word embeddings, contextualized sentence embeddings, and summarization. In addition to raising awareness of these tools, Meyer also empirically evaluates their benefits at improving classification performance.

The symposium’s discussant is Dr. **Chelsea Song**, who is an Assistant Professor of I-O Psychologist at Purdue University and is a recognized expert in applying big data and machine learning methods to study the workplace. Her research seeks to enhance studying of topics related to diversity in the workplace, individual differences (vocational interests, personality), and person-environment fit via applications of machine learning and optimization. Thus, she provides a unique perspective relative to the presenters, but also has expertise that permeates throughout the presentations. Her research has been published in top I-O journals such as the Journal of Applied Psychology, Personnel Psychology, and Journal of Personality and Social Psychology, among others, and has been featured in popular outlets such as Forbes. Dr. Song currently serves on the editorial boards of the Journal of Applied Psychology and Organizational Research Methods.

**References**

Hernandez, I., & Nie, W. (2022). The AI‐IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, Advance online publication.. <https://doi.org/10.1111/peps.12543>

Speer, A. B., Christiansen, N. D., Robie, C., & Jacobs, R. R. (2022). Measurement specificity with modern methods: Using dimensions, facets, and items from personality assessments to predict performance. *Journal of Applied Psychology, 107*(8), 1428–1439. <https://doi.org/10.1037/apl0000618>

Thompson, I., Cubrich, M., Qasim, M., Zhu, Y. E., Hassenkamp, K., Koohifar, F., Koenig, N., & Dudley, N. (2021). A Measurement-based Foundation for AI Applied to the Audio of Interviews. In TMS Proceedings. Technology, Mind, and Society. American Psychological Association. https://doi.org/10.1037/tms0000005

Detecting Careless Responding with Machine Learning Models

Trained by Semi-Synthetic Data

Weiwen Nie

Hogan Assessments

Research consistently confirms a significant proportion of samples often provide careless responses in surveys (Kam & Meyer, 2015). Careless responding inflates item variances, biases item means towards the scale midpoint, and increases residual variances of construct indicators (Goldammer, Annen, Stöckli, & Jonas, 2020). Researchers have developed two types of measures—direct and indirect—to detect careless responses (DeSimone & Harms, 2018). The direct method includes extra items that either provide explicit instruction on how the item should be responded to or bogus items that should be responded to in a specific way. The indirect method focuses on irregularities in the participants’ response patterns. These methods calculate different responding indexes, such as variability psychometric antonyms and psychometric synonyms (Goldammer, et al., 2020). Both types of measures have their criticisms. Critics of the direct methods suggest the items used to measure careless responses can be easily faked (DeSimone & Harms, 2018). Critics of the indirect methods suggest that careless indices are only effective to detect a specific type of careless response (Goldammer et al., 2020) and lower the accuracy of the careless detection method.

To address the criticism of the indirect measure to detect careless response, the current study developed a new indirect careless response detection method based on a machine learning model namely random forest (RF, Breiman, 2001). RF uses a recursive partitioning method to systematically build a set of extensive hierarchical rules and to construct hundreds of decision trees to classify a datapoint’s category (e.g., careless response or not). Each one of the hierarchical rules can reflect a part of an irregular response. For example, for a positive worded and two negative worded items in the same trait, if the person uses the same anchor to respond to these three items, these responses increase the likelihood of being classified as the careless responder. The RF develops hundreds of these hierarchical rules to address many irregular careless response patterns at once.

Two concerns of machine learning models are often discussed are a relatively big amount of data that is required to develop a model and model generalizability across different samples. To address these two concerns, I developed this RF model to detect careless responding with semi-synthetic data. I synthesize the careless response data with a variety of possible careless response patterns and combine it with the non-careless-responses data. This data sythesismethod reduces the need of developing a RF model by half. More importantly, it should enable the performance of the RF model can be generalized to different samples, because the RF model was not developed on a specific sample but the possible careless response pattern.

**Method**

A sample of 800,000 participants who took the Hogan Personality Inventory (HPI, Hogan, 2002) for selection or developmental purposes was used to develop the RF method and the two other indirect careless response detection methods, variability psychometric antonyms and psychometric synonyms methods. The HPI has 4 response anchors ranging from strongly agree to strongly disagree. I split the dataset in half, using random selection to create a control dataset and a manipulated dataset. The manipulated dataset was used to synthesize careless response participants’ samples. I further split the manipulated dataset into 5 smaller manipulated datasets (*n* = 80,000 for each smaller dataset). Within each of the 5 smaller manipulated datasets, a proportion (ranging from 60% - 100% with 10% increasing interval) of items were replaced with randomly generated data. Given no research has identified a clear careless response pattern, the randomly generated data in each smaller dataset was split into 7 parts, and each part used 1 of the 7 approaches (normal distribution, uniform distribution, consecutively choosing a single anchor for 5 to 30 times, only using the middle anchors, only using extreme anchors, only using the lower-end anchors and, only using upper-end anchors) to generate the data

The RF method used an RF model to predict the membership assignment of each participant (control or manipulated). I used a k-folds (k = 3) validation with the random search method to fine-tune the hyperparameters in the RF model.

The “psychometric antonyms” and “psychometric synonyms” methods were developed based on the item level correlation (negative and positive correlations for psychometric antonyms and synonyms methods, respectively) of the control sample (see table 1 for the methods development step).

After these three faking detection methods were developed with the semi-synthetic dataset, they were further validated with a new MTurk sample (N = 875). Participants from the MTurk sample completed the HPI and 12 attention check. Example careless response items included direct instruction item, “Please select agree for this item.”, and bogus item “I eat glass for breakfast this morning.” Participants who failed to respond to any one of the attention check items with the desired anchors were considered careless responders. 683 participants were considered careless responders.

**Results**

Classification accuracy, recall, precision, and F1 score are reported in Table 3. The RF method outperformed the variability psychometric antonyms and psychometric synonyms methods across all classification metrics.

**Discussion**

In the current study, I developed a new indirect careless response detection method based on RF. This RF method outperformed the two traditional indirect methods across all performance metrics. More importantly, the RF method was developed based on a semi-synthetic sample where the careless response was synthesized based on theories and then validated on a new sample in which the participants took the survey for a different purpose. This data synthetization approach addresses two concern in machine learning (ML) research: first, it significantly improves the generalizability of the RF method developed in the current study which is often a weakness of ML-related studies; second, it significantly lowers the need of data requirements of ML models and save practitioners and researchers significant resource and time. Given the encouraging result of the current study, future studies should further explore the possibility of using fully synthesized data to develop the RF careless response detection method to further lower the sample size of non-careless-responding data.

**References**

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

DeSimone, J. A., & Harms, P. D. (2018). Dirty data: The effects of screening respondents

who provide low-quality data in survey research. *Journal of Business and Psychology*,

*33*, 559–577.

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), 101384.

Hogan, R. (2002). The Hogan personality inventory. *Big Five Assessment*, 329-351.

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512-541.

**Table 1.**

*Step-by-Step Variability Psychometric Antonyms and Psychometric Synonyms Methods Development*

|  |
| --- |
| 1. Correlations for all the item pairs were calculated and sorted in ascending order (for psychometric antonyms method) or descending order (for psychometric synonyms method). |
| 2. A range of the item pairs (10 to 150 item pairs with 10 item pairs interval) were selected from the top of both the ascending order and the descending order were select to calculate 140 antonym indexes and 140 synonym indexes respectively. |
| 3. These indexes were calculated by two steps: first, all the item level data were standardized; second, summing up the absolute value of the sum of each selected item pair for each person. |
| 4. A logistic regression model which regressed the member assignment on each of the synonym/antonym index were developed. |
| 5. The logistic regression models that produced the highest accuracy were selected to be the final variability psychometric antonyms and psychometric synonyms methods. This resulted in 80 item pairs for the antonyms method and 70 item pairs for the synonyms method. |

**Table 2**

*Classification Performance of the Variability Psychometric Antonyms method, Variability Psychometric Synonyms Methods and, Random Forest Method*

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Synonym** | **Antonym** | **Random Forest** |
| Accuracy | .79 | .82 | .88 |
| Recall | .82 | .85 | .90 |
| Precision | .89 | .92 | .94 |
| F1 score | .85 | .88 | .92 |

The AI Mirror: Inferring Leader Perceptions Across Demographics

using Natural Language Processing

Fabrice Delice, Ivan Hernandez, & Roseanne Foti

Virginia Tech

Due to cross-cultural differences, leader prototypes (i.e., attitudes, traits, and categories individuals ascribe to a leader) can often vary in how they develop (Gerstner & Day, 1994). Research, thus, falls short in capturing society's collective cognition in leader perception. Moreover, the way leader prototypes manifest is not always a positive reflection upon that society, conferring greater leader status to some groups than others. Considering the importance of studying these divergent attributions and highlighting equity gaps in research, a two-part study was conducted to explore how natural language (NLP) models learn conceptual relationships from a society's discourse. Imbuing with that society's biases, NLP can be applied to understand leader-follower perceptions. These inferences can be leveraged to discover new insights regarding how various social groups are perceived within leader-follower concepts.

**Theoretical Background & Hypotheses**

Previous research has examined the use of leader prototypes, like heuristics and stereotypes, as an effortless and fast way to rate and categorizes others as a leader (Implicit Leadership Theories or *ILT;* Lord et al., 2020) and as a follower (Implicit Followership Theories or *IFT*; Sy, 2012) consciously and unconsciously (Devine, 1989). Leader perception sways expectations and self-standards for expected leader behaviors (Lord et al., 2020).

Social identities are relevant to leader perception, holding social ramifications from career outcomes to the discrimination and mistreatment of members of marginalized groups based on individuals' biases (Cortina, 2008). Exploring social identities within various frameworks such as the Diversity Wheel (Loden & Rosener, 1991) or Wheel of Privilege (Duckworth, 2020) helps us examine discrepancy in perception with a clearer sense of attributes that represents various social categories (Figure 1).

Through self-report methods, exploring general perception faces issues such as underreporting undesirable attitudes or overreporting more desirable ones (i.e., social desirability; Latkin et al., 2017). One solution is Modern NLP which can be applied to derive insights from natural language. Thus, the primary purpose of the current study is to examine the perspectives and opinions related to leader and follower identity found in the associations learned from natural discourse. Prior to this examination, we seek to demonstrate that the associations in these models converge with the mental models for concepts readily addressed with self-report methods. We then will explore biases with the natural language representation to explore perceptions of various identities within leader perception.

**Study 1**

The first study examines hypotheses 1 and 2 (Table 1), which predict that natural language inference (NLI) models produce inferences for concepts related to leader-follower spaces. To demonstrate this convergence, we use archival data collected from (Roediger et al., 2016) that examined people’s perceptions of how various traits are associated with a leader and follower. We then compared the mental map created by NLI models for those terms to the mental map derived from human ratings.

**Method**

In Roediger et al., (2016), participants filled out an online survey that included a modified Q-sort task. Each participant sorted a set of 119 characteristics into five mutually exclusive and predetermined categories: 1) "Leader" 2) “Follower” 3) "A Leader as well as a Follower" 4) “Fits Neither Group” and 5) “I am unsure of the word’s meaning.” The 119 characteristics were obtained from prominent papers on ILT and IFT. To draw inferences regarding the association between leader traits, we employed the Adversarial XLNet NLI models (Nie et al., 2020), trained to predict the probability of a hypothesis statement being true, unknowable, or false given a premise statement (Figure 2). We then calculated the implied association between attributes through pairwise combinations for each trait through a 119 x 119 symmetric dissimilarity matrix by assigning one trait in the premise, "That person is [trait 1]," and then specifying the other trait as the hypothesis term in the model, "That person is [trait 2]. We visualized the associations between concepts in two-dimensions using Multidimensional Scaling on the dissimilarity matrix.

**Results & Discussion**

Study 1 supported hypotheses 1 and 2 by demonstrating convergence between the conceptual associations of leader-follower relevant traits that humans produce, and the associations produced by NLI models. The correlation between the distances (*r*=0.75, *p*<.001.) exceeded conventional reliability thresholds (*r*=.70). In addition to showing reliability-like correlations between human and AI inferred trait associations (Figure 3), we also found that the terms followed the theoretical octant structure of personal traits (Wiggins, 1979; Figure 4). When mapping properties of leaders and followers within that space, NLI correctly identifies the directions within the space that correspond to how much a leader, follower, both, and neither is (Figure 5).

**Study 2**

Study 2’s goal is to identify whether specific social identities that are high or low in

privilege is aligned more closely with leader dimensions by the natural language model.

**Methods**

Using Black & Stone's (2005) study, we identified social identities from groups that hold power/privilege to those with relatively less (Table 2). We rotated the perceptual map so that the dimension that uniquely separated leaders and followers, was aligned with the x-axis. Therefore, the axis becomes a unique “leader-follower” dimension. We then fit a univariate regression using the inferred probability of the term given the group member status as the outcome, and the x-axis coordinates of each term (i.e., the “unique leader” dimension). The resulting coefficient represents the relative position of the subgroup on the leader dimension (Figure 6).

**Results**

Study 2 found support for hypothesis 3 (Table 1) by illustrating that, within natural discourse, group members within social identities that have lower privilege have more “follower-like” patterns of associations with leader-follower traits. These patterns were found for SES, disability, citizen status, education, geographic environment, gender roles (Table 3 -8).

**General Discussion & Conclusion**

Our results supported NLI models' ability to provide convergent and novel insights into collective leader-follower perceptions. The orientation discovered by multidimensional scaling inferred similarities between leader traits, corresponding to prior research on leaders but also on interpersonal trait perception. These embedded perceptions can illustrate the positive and negative aspects of that society, identifying disparities in experiences by groups that may lack the privilege or voice to express them.

**References**

Black, L. L., & Stone, D. (2005). Expanding the definition of privilege: The concept of social privilege. *Journal of Multicultural Counseling and Development*, *33*(4), 243-255. [https://doi.org/10.1002/j.2161-1912.2005.tb00020.x](https://psycnet.apa.org/doi/10.1002/j.2161-1912.2005.tb00020.x)

Cortina, L. M. (2008). Unseen injustice: Incivility as modern discrimination in organizations. *Academy of Management Review, 33*(1), 55–75.<https://doi.org/10.5465/amr.2008.27745097>

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. [https://doi.org/10.1037/0022-3514.56.1.5](https://psycnet.apa.org/doi/10.1037/0022-3514.56.1.5)

Duckworth, S. (2020, Oct 18). Wheel of power/privilege [Infographic]. Flickr. https://flic.kr/p/2jWxeGG. CC BY-NC-ND 2.0

Gerstner, C. R., & Day, D. V. (1994). Cross-cultural comparison of leadership prototypes. *The Leadership Quarterly*, *5*(2), 121-134. [https://doi.org/10.1016/1048-9843(94)90024-8](https://psycnet.apa.org/doi/10.1016/1048-9843(94)90024-8)

Latkin, C. A., Edwards, C., Davey-Rothwell, M. A., & Tobin, K. E. (2017). The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland. *Addictive Behaviors*, *73*, 133-136. [https://doi.org/10.1016/j.addbeh.2017.05.005](https://psycnet.apa.org/doi/10.1016/j.leaqua.2014.09.002)

Loden, M., & Rosener, J. B. (1991). *Workforce America!: Managing employee diversity as a vital resource*. McGraw-Hill.

Lord, R. G., Epitropaki, O., Foti, R. J., & Hansbrough, T. K. (2020). Implicit leadership theories, implicit followership theories, and dynamic processing of leadership information. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*, 49-74.<https://doi.org/10.1146/annurev-orgpsych-012119-045434>

Roediger, M., Coyle, P., Shah, Y., Poli, R. A., Burns, D., & Foti, R. J. (April, 2016). Using correspondence analysis to measure implicit leader and follower theories. In P. Coyle & R. J. Foti (Co-chairs), *Measuring Leadership and Followership: Clarifying constructs and items.* Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology, Anaheim, CA.

Sy, T. (2010). What do you think of followers? Examining the content, structure, and consequences of implicit followership theories. *Organizational Behavior and Human Decision Processes*, *113*(2), 73-84. <https://doi.org/10.1016/j.obhdp.2010.06.001>

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology, 37*(3), 395–412. <https://doi.org/10.1037/0022-3514.37.3.395>

**Table 1**

*Summary of Study 1 & Study 2 Hypotheses*

|  |  |  |
| --- | --- | --- |
| **Study** | **Hypotheses** | **Findings** |
| Study 1 | *Hypothesis 1:* Human perceptions of leader and follower trait interrelations will converge with the interrelations suggested by natural language inferences. | Supported |
|  | *Hypothesis 2:* Human perceptions of conceptual groupings will converge with the groupings suggested by natural language inferences. | Supported |
| Study 2 | *Hypothesis 3a:* Members of higher privileged groups are seen as more reflective of leader traits | Supported |
|  | *Hypothesis 3b:* Members of lower privileged groups are seen as more reflective of follower traits | Supported |

**Table 2**

*Black & Stone's (2005) Social identities Power/Privilege vs. Marginalized*

|  |  |  |
| --- | --- | --- |
| **Identities** | **Power/Privilege** | **Marginalized** |
| Socioeconomic Status | Rich | Poor |
| Citizenship Status | U.S. Citizen | Undocumented |
| Disability Status | Able-bodied | Disabled |
| Environment | Urban | Rural |
| Gender Roles | Masculinity | Femininity |
| Education | Educated | Uneducated |

*Note.* The following social identities were compared in the study. We were interested in comparing leader perceptions between groups of higher privilege and their more marginalized counterparts. Although prior taxonomies identify a variety of privileged/marginalized identities, to provide a more narrowed focus of the investigation, we combined constructs that are conceptually similar (i.e., race and ethnicity) and did not investigate highly multifaceted concepts (i.e., “appearance”, “interests”), emphasizing concepts with clear high privilege and low privilege categories.

**Table 3**

*Regression Table for Socioeconomic Status Model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *B* | *β* | *t* | 95*%* *CI* | | *p-value* |
|  |  |  |  | LL | UL |  |
| Rich | 0.353 | 0.359 | 4.998 | 0.216 | 0.501 | < 0.001 |
| Poor | -0.584 | -0.413 | -5.11 | -0.574 | -0.253 | < 0.001 |
| Rich X Poor | 0.919 | 0.696 | 6.160 | 0.475 | 0.917 | < 0.001 |

*Note.* The position of high SES people is higher on the unique leader dimension (*B* = 0.353, *β* = 0.359, *t*(119) = 4.998, *p* < 0.001, 95*%* *CI* =[0.216, 0.501]) compared to their low SES counterparts (*B* = -0.584, *β* = -0.413, *t*(119) = -5.11, *p* < 0.001, 95*%* *CI* =[-0.574, -0.253]). While high SES people were perceived on the positive side of the unique leader dimension, the negative coefficient for low SES people suggests a closer association to submissiveness. The interaction test between coefficients found that the difference was statistically significant (*B* = 0.919, *β* = 0.696, *t*(119) = 6.160, *p* < 0.001, 95*%* *CI* =[0.475, 0.917]).

**Table 4**

*Regression Table for Citizenship Status Model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *B* | *β* | *t* | 95*%* *CI* | | *p-value* |
|  |  |  |  | LL | UL |  |
| Citizen | 0.011 | 0.011 | 0.135 | -0.147 | 0.169 | 0.893 |
| Undocumented | -0.166 | -0.126 | -1.625 | -0.280 | 0.028 | 0.107 |
| Citizen X Undocumented | 0.153 | 0.127 | 0.986 | -0.125 | 0.378 | 0.324 |

*Note.* The position of U.S. citizens is higher on the unique leader dimension (*B* = 0.011, *β* = 0.011, *t*(119) = 0.135, *p* = 0.893, *95% CI* =[-0.147, 0.169]) compared to undocumented people (*B* = -0.166, *β* = -0.126, *t*(119) = -1.625, *p* = 0.107, *95% CI* =[-0.280, 0.028]). U.S. citizens were perceived on the positive side of the unique leader dimension while undocumented people’s negative coefficient indicates a closer association with submissiveness. However, the interaction test between coefficients found that the difference was not statistically significant (*B* = 0.153, *β* = 0.127, *t*(119) = 0.986, *p* = 0.324, *95% CI* =[-0.125, 0.378]).

**Table 5**

*Regression Table for Disability Status Model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *B* | *β* | *t* | 95*%* *CI* | | *p-value* |
|  |  |  |  | LL | UL |  |
| Able-bodied | 0.233 | 0.235 | 3.001 | 0.080 | 0.391 | 0.003 |
| Disabled | -0.709 | -0.500 | -7.087 | -0.640 | -0.360 | < 0.001 |
| Able-bodied X Disabled | 0.922 | 0.585 | 6.299 | 0.403 | 0.767 | < 0.001 |

*Note*. The position of able-bodied is higher on the unique leader dimension (*B* = 0.233, *β* = 0.235, *t*(119) = 3.001, *p* = 0.003, *95% CI* =[0.080, 0.391]) compared to disabled people (*B* = -0.709, *β* = -0.500, *t*(119) = -7.087, *p* < 0.001, *95% CI* =[-0.640, -0.360]). Able-bodied people were perceived on the positive side of the unique leader dimension, while disabled people’s negative coefficient indicates a closer association with submissiveness (Figure 10). As of the interaction test for the difference between coefficients, the difference in coefficients is statistically significant (*B* = 0.922, *β* = 0.585, *t*(119) = 6.299, *p* < 0.001, *95% CI* =[0.403, 0.767]).

**Table 6**

*Regression Table for* Environment *Model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *B* | *β* | *t* | 95*%* *CI* | | *p-value* |
|  |  |  |  | LL | UL |  |
| Urban | 0.253 | 0.391 | 4.728 | 0.227 | 0.555 | < 0.001 |
| Rural | -0.049 | -0.092 | 0.983 | 0.278 | 0.094 | 0.328 |
| Urban X Rural | 0.300 | 0.505 | 4.292 | 0.275 | 0.736 | < 0.001 |

*Note.* The position of urban is higher on the unique leader dimension (*B* = 0.253, *β* = 0.391, *t*(119) = 4.728, *p* < 0.001, *95% CI* =[0.227, 0.555]) compared to rural people (*B* = -0.049, *β* = -0.092, *t*(119) = -0.983, *p* = 0.328, *95% CI* =[-0.278, 0.094]). Urban people were perceived on the positive side of the unique leader dimension, while rural people’s negative coefficient indicates a closer association with submissiveness. Results from the interaction test which examined the difference between coefficients, there was a statistically significant difference (*B* = 0.300, *β* = 0.505, *t*(119) = 4.292, *p* < 0.001, *95% CI* =[0.275, 0.736]).

**Table 7**

*Regression Table for Education Status Model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *B* | *β* | *t* | 95*%* *CI* | | *p-value* |
|  |  |  |  | LL | UL |  |
| Educated | 0.446 | 0.285 | 4.410 | 0.157 | 0.413 | < 0.001 |
| Uneducated | -0.444 | -0.272 | -3.886 | -0.410 | -0.133 | < 0.001 |
| Able-bodied X Disabled | 0.852 | 0.526 | 4.213 | 0.281 | 0.770 | < 0.001 |

*Note.* The position of educated is higher on the unique leader dimension (*B* = 0.446, *β* = 0.285, *t*(119) = 4.410, *p* < 0.001, *95% CI* =[0.157, 0.413]) compared to uneducated people (*B* = -0.444, *β* = -0.272, *t*(119) = -3.886, *p* < 0.001, *95% CI* =[-0.410, -0.133]). Educated people were perceived on the positive side of the unique leader dimension, while uneducated people’s negative coefficient indicates a closer association with submissiveness (Figure 12). The interaction test showed a statistically significant difference between coefficients (*B* = 0.852, *β* = 0.526, *t*(119) = 4.213, *p* < 0.001, *95% CI* =[0.281, 0.770]).

**Table 8**

*Regression Table for Gender Roles Model*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *B* | *β* | *t* | 95*%* *CI* | | *p-value* |
|  |  |  |  | LL | UL |  |
| Rich | 0.790 | 0.702 | 11.202 | 0.578 | 0.826 | < 0.001 |
| Poor | -0.438 | -0.333 | -4.982 | -0.465 | -0.200 | < 0.001 |
| Rich X Poor | 1.237 | 1.000 | 11.680 | 0.832 | 1.167 | < 0.001 |

*Note.* The position of masculine is higher on the unique leader dimension (*B* = 0.790, *β* = 0.702, *t*(119) = 11.202, *p* < 0.001, *95% CI* =[0.578, 0.826]) compared to feminine people (*B* = -0.438, *β* = -0.333, *t*(119) = -4.982, *p* < 0.001, *95% CI* =[-0.465, -0.200]). Both masculine and feminine people were perceived on the positive side of the unique leader dimension, which may reflect that feminine attributes are not inherently perceived as followers, but rather may simply be a different type of leader prototype. In terms of the interaction test, the results indicated a statistically significant difference between coefficients (*B* = 1.237, β = 1.000, *t*(119) = 11.680, *p* < 0.001, 95*%* *CI* =[0.832, 1.167]).

**Figure 1.**

*Diversity Wheel and Wheel of Privilege*

Engineering drawing

Description automatically generatedChart, sunburst chart

Description automatically generated

*Note.* The left panel shows how there are various social identities that comprise diversity. The right panel illustrates where privilege is found within various subgroups of those identities.

**Figure 2.**

*Illustration of Natural Language Inference Model*

Shape

Description automatically generated with medium confidence

*Note.* The above figure shows how a natural language inference model accepts a premise statement and a hypothesis statement. It then scores how well the hypothesis can be inferred, given the premise. The panel on the right provides a concrete example of how the model would score traits related to person perception. We subtracted the contradiction score from the entailment score to obtain a single value of association between traits.

**Figure 3.**

*Comparison of Original vs Inferred Perceptual Map of Leader, Follower, Both, Neither terms*

*Scatter chart

Description automatically generated*Text

Description automatically generated with medium confidence

***Note.*** *The left panel contains the perceptual map constructed from human perceptions. The right panel contains the perceptual map derived from the natural language inference model. The color of the data point indicates the cluster the term was assigned by the authors from the original human-based study. blue = leader, violet = leader and both, purple = both, magenta = follower and both, red = follower, brown = follower and neither, green = neither, turquoise = leader and neither.*

**Figure 4.**

*Comparison of the Interpersonal Circumplex and the Natural Language Model Inferred Projections of Each Dimension within Perceptual Space*

A picture containing text, receipt

Description automatically generatedDiagram

Description automatically generated with medium confidence

*Note.* The left panel contains the theoretical alignment of the eight dimensions (octants) of interpersonal traits (Wiggins 1979). The right panel contains the perceptual map derived from the natural language inference model. The projection vectors are the regression coefficients from the property fitting regression. The property fitting regression uses the natural language model’s inferred entailment of each trait, given the octant-related traits, as the outcome variable. The x and y coordinates of each trait are the predictor variables.

**Figure 5.**

*Comparison of Original vs Inferred Property Fitting of Leader, Follower, Both, and Neither*

*within the Perceptual Space.*

Chart, scatter chart

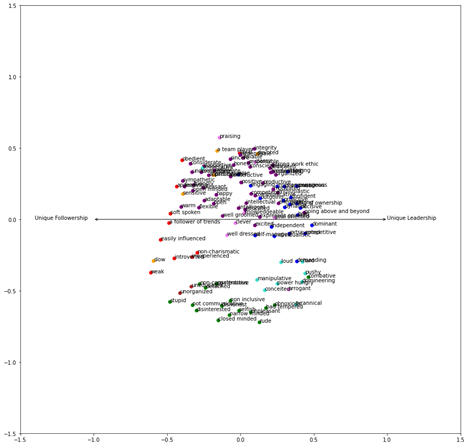
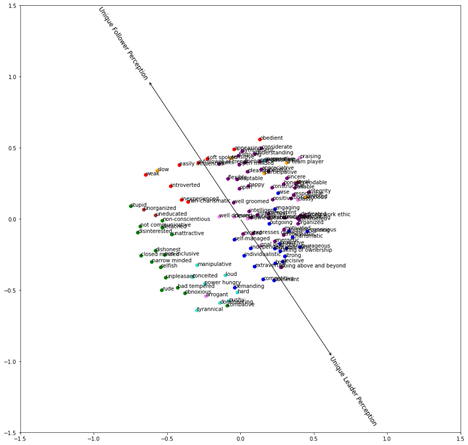
Description automatically generatedChart, scatter chart

Description automatically generated

*Note.* The left panel contains the perceptual map constructed from human perceptions. The right panel contains the perceptual map derived from the natural language inference model. Lines have been rescaled to fit the chart. The vectors on the left panel were computed using correspondence analysis from human data, and the vectors on the right are the regression coefficients of the natural language inference model’s entailment prediction of each trait, given the property (e.g., “I am uniquely a leader”) using the x and y coordinate of each trait at the predictor variables.

***Figure 6***

*Demonstration of how Perceptual Map was Rotated to Capture “Unique Leader” on X-axis*

******

*Note.* The left panel shows a line where unique-leadership traits are highest vs lowest based on fit property vectors. We rotated the figure so that the unique-leadership is the x-axis. This rotation allows predicting a social identity’s association between traits from a natural language inference model using its position in space (x,y coordinates). Higher/larger slopes for the x coordinate, indicate higher perceptions along the unique-leadership dimension

Leadership Construct Consolidation via NLP: Identifying the Overlap between Conceptually Related Items

Bryan Acton, Mengying Li, & Chunghyun Oh

Binghamton University

Follower-reported ratings of leaders represent the core method used in the leadership literature. This methodology has thus shaped both leadership theory and practice: driving both the development of new theories, as well as organizational decisions (e.g., promotions). Unfortunately, throughout the history of leadership research, significant concern has been amassed regarding these measurement techniques (e.g., Hansbrough et al., 2015; Hunter et al., 2007). In particular, despite the leadership literature developing large and distinct sub-literatures built around different leadership constructs (e.g., authentic, transformational), there are questions as to whether these constructs are actually unique from one another (e.g., van Knippenberg & Sitkin, 2013; Yammarino et al., 2020). While there is past work testing the redundancy across particular leadership constructs (e.g., authentic vs. transformational; Banks et al., 2016), it is difficult to test these claims across the entire landscape of leadership constructs due to its large scale. In the current paper, we argue that big data text analysis methods (i.e., Natural Language Processing) can be used to create a comprehensive test of the redundancy across different prominent leadership constructs. Specifically, in utilizing a recent big-data algorithm (1 million items) that can accurately assess the expected similarity across psychometric items (Hernandez & Nie, in press), we test for the conceptual overlap among prominent leadership scales.

**Leadership Scale Measurement**

In their review of the entire leadership literature, Dinh and colleagues (2014) found over 60 unique types of leadership theories that each have unique associated constructs and measurement scales. Charismatic leadership, Servant leadership, and Transformational leadership represent examples of these unique constructs. Similar to how personality has unique dimensions, leadership has different scales with unique items for each construct. For example, “My leader communicates a clear and positive vision of the future” is an example of an item from a Transformational Leadership Scale. Meuser and colleagues (2016) expanded Dinh and colleagues’ work by using a network approach to find that each leadership theory was rarely integrated with other theories. From this work and others (Day & Antonakis, 2013), it is clear that leadership (1) relies on constructs represented by follower-reported questionnaires (Hansbrough et al, 2015), and (2) these constructs are treated as distinct from one another. Based on these ideas, we would expect that (1) leadership items developed for each respective construct would more similarly relate to items from the same construct, and (2) items would remain distinct from items that derive from different constructs. The purpose of the current work is to test these ideas using a Natural Language Processing Algorithm that is trained to predict the co-relatedness of psychometric items.

**Methods**

We used a text-based machine learning approach to examine the semantic relatedness among leadership items. Specifically, we relied on an SBERT model, which is a transformer model that creates sentence-text embeddings– a latent representation of semantic meaning (Reimers & Gurevych, 2019). This model takes text from leadership items as an input, and then provides the expected similarity with other leadership items as an output. To train a model on semantic similarity, we relied on the recent work of Hernandez & Nie (2020), which utilized thousands of psychometric scale items to train a model that accurately predicts the expected covariation between items. Next, in order to update the model to account for which construct the item belongs to, we relied on a triple extraction technique (e.g., Deng & Liu, 2021). Items that derive from the same construct received a positive weight, and items from different constructs received a negative weight. We trained the model using items from the IPIP personality item database (N = 3805). After training the model, we held out 10 percent of the sample as a cross-validation sample. The final model surpassed 95% accuracy at creating embeddings that were similar for items that derive from the same construct. The full model building process is illustrated in Figure 1.

**Results**

To test the model on leadership items, we first selected scales from three dominant leadership constructs: charismatic, servant, and transformational.[[1]](#footnote-1) As the semantic similarity embeddings represent a multidimensional space, we utilized a dimension reduction technique to visualize the similarities between semantic embeddings. To illustrate the relatedness between items from different constructs, we selected items from five distinct constructs from our personality validation sample. As displayed in Figure 2, the difference between clusters shows how related they are. There are five distinct clusters in Figure 1, showing that items tended to be most similar to items from the same construct.

Next, we performed the same analysis on the leadership items from the three different constructs. The results are displayed in Figure 3. Overall, there appear to be about five clusters (Figure 3). However, these clusters are made up of various items from the three different leadership scales (Figure 3). These results provide evidence that these constructs may not be distinct from one another and that items from one construct may be contaminated with those from another construct.

**Discussion**

Overall, our approach demonstrates a possible method that can address the question of construct redundancy and contamination in leadership measurement. The advantage that this method holds over other methods is that it does not require the collection of new ratings data to assess the similarities among items. While analyzing particular scales has proven successful (e.g., Banks et al., 2016), this can be resource intensive. Our approach supports earlier work, but presents the opportunity to gather evidence across the full spectrum of leadership scales (i.e. Dinh et al., 2014).

As the results show that items tend to not cluster around constructs, what determines their similarity? We argue that the work of Lord and colleagues (2021) may provide important insights. They argue that most leadership ratings are based on core attributes such as (1) abstractness/concreteness, (2) emotionality, and self-reference. While we plan to further investigate the role of these attributes, the current results suggest that leadership measurement does have a construct contamination problem.

**References**

Banks, G. C., McCauley, K. D., Gardner, W. L., & Guler, C. E. (2016). A meta-analytic review of authentic and transformational leadership: A test for redundancy. *The Leadership Quarterly, 27(4)*, 634–652. https://doi.org/10.1016/j.leaqua.2016.02.006

Day, D. V., & Antonakis, J. (2013). The future of leadership. In The Wiley-Blackwell handbook of the psychology of leadership, change, and organizational development (pp. 221–235). Wiley Blackwell. https://doi.org/10.1002/9781118326404.ch11

Deng, W., & Liu, Y. (2021). Chinese Triple Extraction Based on BERT Model. 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), 1–5. https://doi.org/10.1109/IMCOM51814.2021.9377404

Dinh, J. E., Lord, R. G., Gardner, W. L., Meuser, J. D., Liden, R. C., & Hu, J. (2014). Leadership theory and research in the new millennium: Current theoretical trends and changing perspectives. *The Leadership Quarterly, 25(1)*, 36–62. https://doi.org/10.1016/j.leaqua.2013.11.005

Hansbrough, T. K., Lord, R. G., & Schyns, B. (2015). Reconsidering the accuracy of follower leadership ratings. *The Leadership Quarterly, 26(2)*, 220–237. https://doi.org/10.1016/j.leaqua.2014.11.006

Hernandez, I., & Nie, W. (in press). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*. https://doi.org/10.1111/peps.12543

Hunter, S. T., Bedell-Avers, K. E., & Mumford, M. D. (2007). The typical leadership study: Assumptions, implications, and potential remedies. *The Leadership Quarterly, 18(5)*, 435–446. https://doi.org/10.1016/j.leaqua.2007.07.001

Lord, R. G., Epitropaki, O., Foti, R. J., & Hansbrough, T. K. (2020). Implicit Leadership Theories, Implicit Followership Theories, and Dynamic Processing of Leadership Information. A*nnual Review of Organizational Psychology and Organizational Behavior, 7(1)*, 49–74. https://doi.org/10.1146/annurev-orgpsych-012119-045434

Meuser, J. D., Gardner, W. L., Dinh, J. E., Hu, J., Liden, R. C., & Lord, R. G. (2016). A Network Analysis of Leadership Theory: The Infancy of Integration. *Journal of Management, 42(5)*, 1374–1403. https://doi.org/10.1177/0149206316647099

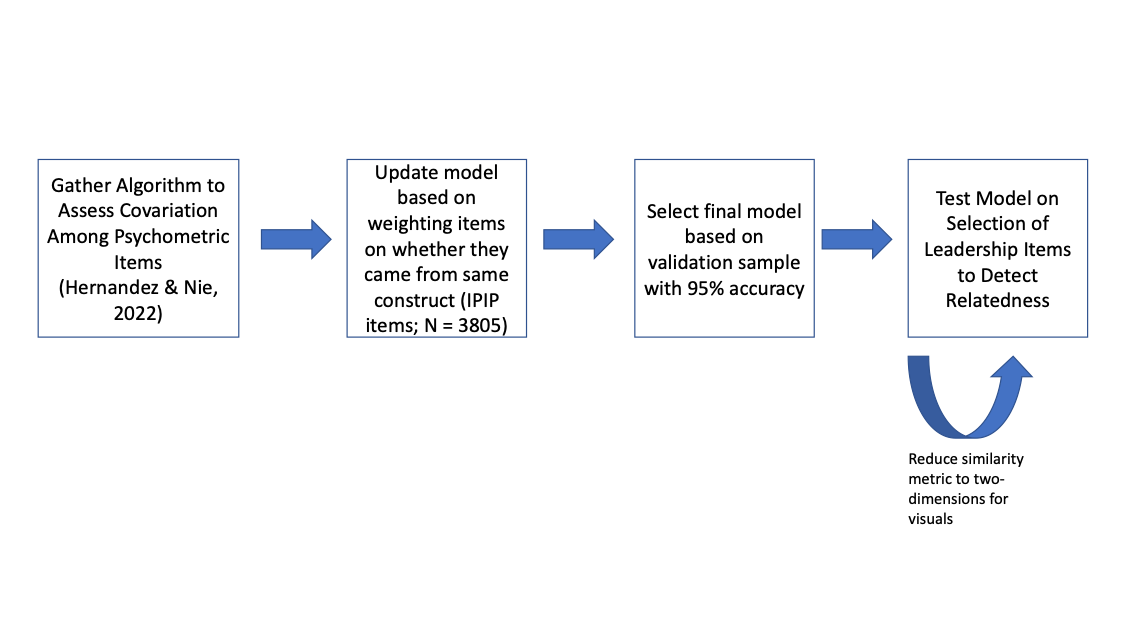
Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. https://doi.org/10.18653/v1/D19-1410

van Knippenberg, D., & Sitkin, S. B. (2013). A Critical Assessment of Charismatic—Transformational Leadership Research: Back to the Drawing Board? *Academy of Management Annals, 7(1)*, 1–60. https://doi.org/10.5465/19416520.2013.759433

Yammarino, F. J., Cheong, M., Kim, J., & Tsai, C.-Y. (2020). Is Leadership More Than “I Like My Boss”? In M. Ronald Buckley, A. R. Wheeler, J. E. Baur, & J. R. B. Halbesleben (Eds.), Research in Personnel and Human Resources Management (Vol. 38, pp. 1–55). Emerald Publishing Limited. https://doi.org/10.1108/S0742-730120200000038003

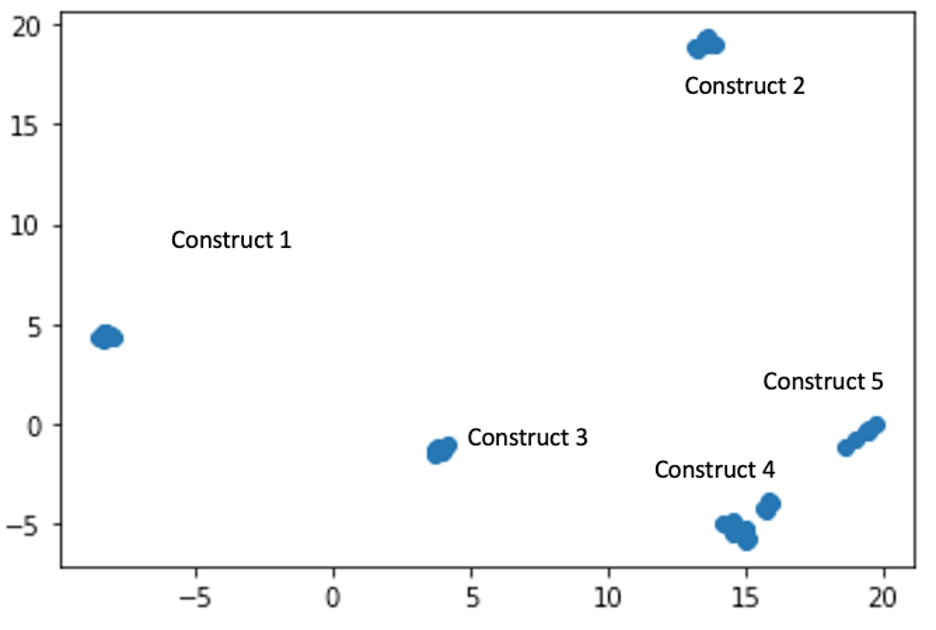
**Figure 1.**

*Natural Language Development Process to Detect Relatedness*



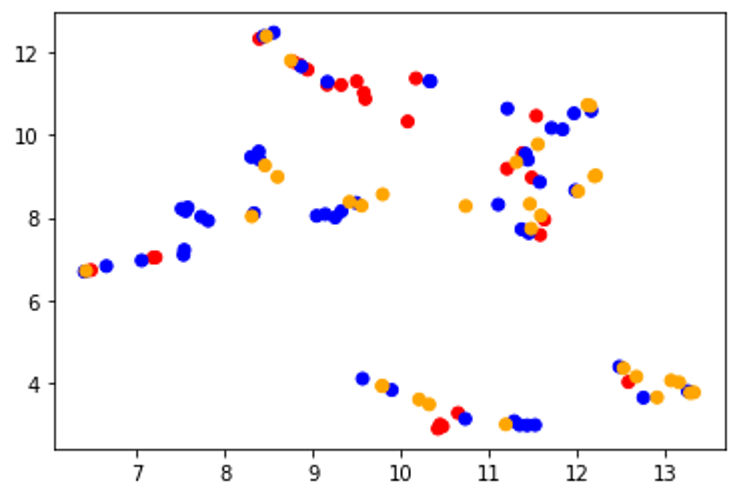
**Figure 2.**

*Embedding similarity among five distinct personality constructs*



**Figure 3.**

*Embedding similarity across three prominent leadership scales*

****

*Notes. Red = charismatic leadership. Blue = servant leadership. Orange = Transformational leadership.*

Experimenting with the Dark Arts of Data Augmentation for Natural Language Processing

Joe Meyer1, Nick Koenig2

*Meta*1*, Walmart*2

Natural Language Processing (NLP) and Machine Learning (ML) have revolutionized many different fields and technologies, for example smart assistants (e.g., Siri and Alexa), search (e.g., BERT; Devlin et al., 2019), and identifying harmful content (e.g., Metzler, Baginski, Niederkrotenthaler, & Garcia, 2022). Additionally, the transfer learning paradigm (Howard, & Ruder, 2018) have allowed data hungry deep neural network architectures to generalize across diverse tasks through rigorous pre-training and fine-tuning on local data, for example employing masked-language modeling and next sentence prediction (Devlin et al., 2019) and fine-tuning on a linguistic acceptability task (CoLA; Warstadt, Singh,, & Bowman, 2019). However, gathering local, task-specific data is still nonetheless resource intensive. This paper presents several data augmentation methods for NLP and compares their performance on an open-source dataset. Data augmentation manipulates the model training process and has been applied in many subfields of Artificial Intelligence (AI), for example computer vision (Krizhevsky, Sutskever, & Hinton, 2017). To the best of our knowledge, no papers have been published on data augmentation for Natural Language Processing in Industrial-Organizational (I-O) Psychology. This study outlines several data augmentation methods including synonym replacement, contextualized word embeddings, language generation, and summarization. We present initial results and additionally, we present a method to preserve data augmentation quality.

**Method**

We chose the widely used IMDB Sentiment dataset (Maas et al., 2011) for this experiment. The data contains human annotation of sentiment of sentences. The dataset was downloaded directly from Hugging Face. The data (n=50000) was randomly sampled (n=7500) for the initial experiment. The data was split into a train(n =7200), validation (n = 800), and test (n = 2000) set. We randomly sampled 5%(n = 540) of the training dataset which was used for data augmentation. The same data was augmented for each data augmentation technique. Augmentation was performed in Google Colab, while model training was performed on a larger remote GPU.

**Synonyms**

In synonym replacement, n words are selected from the document. These n words are then replaced with words with similar meanings. We use the wordnet dictionary (Miller, 1995), which is a large database of semantic relationships between English words, to replace words with synonyms. For the initial model, a minimum of one word is replaced with a maximum of five words.

**Contextualized Word Embeddings**

Word embeddings (Mikolov, Chen, Corrado, & Dean, 2013) are dense vector representations of words trained via self-supervised neural networks that represent word meaning. However, some word embeddings are static, and have the same representation across different use cases. Contextualized word embeddings take into account word context by leveraging techniques such as positional embeddings. During data augmentation rather than simple synonym replacement, the word’s context is taken into account when determining a high-quality substitution. We use the bert-base-uncased model to substitute words. The same default settings are used, where there is a minimum of 1 word being replaced and a maximum of five words replaced.

**Language Generation**

We use the GPT-2 (Radford et al., 2019) model that is capable of producing nearly human-generated language to expand on training data by using the original texts as seeds to produce more language. Settings included a minimum length of 10 and a maximum length of 100. Additionally, top\_p is set to .9, which thresholds the likelihood of tokens that are selected.

**Abstractive Summarization**

One task of natural language processing is being given a text and automatically producing a summary that encapsulates the meaning of the original text. We use the t5-small model (Raffel et al., 2020) with the default settings of a minimum length of 20 and a maximum length of 50. Additionally, batch size was 32 and top\_p was .9.

**Model Training**

After data augmentation is performed, the augmented data sets are concatenated with the original training set. Additionally, the original training set is used as a baseline. We use the BERT-base (Devlin et al., 2019) model to train supervised learning classifiers on each of the datasets (n=6). Models are evaluated on the validation dataset every 1000 training steps. We train all models for four epochs with a batch size of 16.

**Results**

Initial results indicate that data augmentation marginally increases performance beyond that baseline. Synonym, contextualized word embeddings, and abstractive summarization all had an F1 score of .92 on the hold out set (n=2000), while the baseline and language generation had a .91 F1 score. More analyses will be performed to gather a more in-depth perspective on the effect of data augmentation.

**Future Directions**

Next, we plan to optimize hyperparameters for each of the data augmentation models. For example, in synonym replacement a minimum of one word is changed, with a maximum of five. It’s possible that optimizing the number of substitutions can improve modeling performance. Similarly, the initial results are from default settings for each augmenter model. Additionally, we augmented a relatively small portion of the training data (5%). Data augmentation’s effects may be more pronounced when the ratio of original data to augmented data is smaller. Also, we plan to examine the effect of model choice on performance as well, for example using Distill-BERT (Sanh, Debut, Chaumond, & Wolf, 2019) and RoBERTa-base (Liu et al., 2019) models. Finally, we plan to create a customized augmented data quality model that preserves the original meaning of the data.

***Augmented Data Quality Filter***

We plan to use cosine similarity as a decision making function whether or not to retain an augmented text. Cosine similarity is a normalized form of the dot product that computes the degree of similarity between two vectors. If two vectors have a high degree of similarity, then they are close in semantic meaning. We plan to vectorize the original texts and the augmented texts using Sentence BERT (Reimers, & Gurevych, 2019) to create dense embeddings, then compute the cosine similarity between them for all original-augmented pairs. Then, we plan to set *n* thresholds *s* between 0 and 1 and examine the effect of retaining augmented text - original text similarity > *s* on accuracy.

**References**

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*.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142-150).

Metzler, H., Baginski, H., Niederkrotenthaler, T., & Garcia, D. (2022). Detecting potentially harmful and protective suicide-related content on Twitter: machine learning approach. *Journal of Medical Internet Research*, *24*(8), e34705.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(140), 1-67.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Cola: The corpus of linguistic acceptability (with added annotations).

**Table 1.**

*Classification Performance of Various Data Augmentation Approaches*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | 0 | 1 | accuracy | macro avg | weighted avg | Model |
| precision | 0.90 | 0.95 | 0.92 | 0.93 | 0.93 | Synonym\_Aug |
| recall | 0.95 | 0.90 | 0.92 | 0.92 | 0.92 | Synonym\_Aug |
| f1-score | 0.93 | 0.92 | 0.92 | 0.92 | 0.92 | Synonym\_Aug |
| support | 994.00 | 1006.00 | 0.92 | 2000.00 | 2000.00 | Synonym\_Aug |
| precision | 0.93 | 0.90 | 0.92 | 0.92 | 0.92 | ContextualWordEmbs\_Aug |
| recall | 0.90 | 0.94 | 0.92 | 0.92 | 0.92 | ContextualWordEmbs\_Aug |
| f1-score | 0.91 | 0.92 | 0.92 | 0.92 | 0.92 | ContextualWordEmbs\_Aug |
| support | 994.00 | 1006.00 | 0.92 | 2000.00 | 2000.00 | ContextualWordEmbs\_Aug |
| precision | 0.88 | 0.95 | 0.91 | 0.92 | 0.92 | ContextualWordEmbsForSentence\_Aug |
| recall | 0.95 | 0.88 | 0.91 | 0.91 | 0.91 | ContextualWordEmbsForSentence\_Aug |
| f1-score | 0.92 | 0.91 | 0.91 | 0.91 | 0.91 | ContextualWordEmbsForSentence\_Aug |
| support | 994.00 | 1006.00 | 0.91 | 2000.00 | 2000.00 | ContextualWordEmbsForSentence\_Aug |
| precision | 0.90 | 0.93 | 0.92 | 0.92 | 0.92 | AbstSumm\_Aug |
| recall | 0.94 | 0.90 | 0.92 | 0.92 | 0.92 | AbstSumm\_Aug |
| f1-score | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | AbstSumm\_Aug |
| support | 994.00 | 1006.00 | 0.92 | 2000.00 | 2000.00 | AbstSumm\_Aug |
| precision | 0.93 | 0.90 | 0.91 | 0.91 | 0.91 | baseline |
| recall | 0.89 | 0.94 | 0.91 | 0.91 | 0.91 | baseline |
| f1-score | 0.91 | 0.92 | 0.91 | 0.91 | 0.91 | baseline |
| support | 994.00 | 1006.00 | 0.91 | 2000.00 | 2000.00 | baseline |

1. We are in the process of testing this model on a much larger database of leadership items. We will present the full results at the conference. You can see results from for larger sample of items (N = 536) at https://github.com/actonbp/NLP\_Leader\_items [↑](#footnote-ref-1)