한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2021. Vol. 34, No. 2, 213-236
https://doi.org/10.24230/kjiop.v34i2.213-236

Machine Learning Approach to Personality Assessment and Its Application to Personnel Selection: A Brief Review of the Current Research and Suggestions for the Future

Jisoo Ock† Hyeryeon An‡

Department of Business Administration, Pusan National University

As we enter the digital age, new methods of personality testing-namely, machine learning-based personality assessment scales-are quickly gaining attraction. Because machine learning-based personality assessments are made based on algorithms that analyze digital footprints of people's online behaviors, they are supposedly less prone to human biases or cognitive fallacies that are often cited as limitations of traditional personality tests. As a result, machine learning-based assessment tools are becoming increasingly popular in operational settings across the globe with the anticipation that they can effectively overcome the limitations of traditional personality testing. However, the provision of scientific evidence regarding the validity psychometric soundness and the fairness of machine learning-based assessment tools have lagged behind their use in practice. The current paper provides a brief review of empirical studies that have examined the validity of machine learning-based personality assessment, focusing primarily on social media text mining method. Based on this review, we offer some suggestions about future research directions, particularly regarding the important and immediate need to examine the machine learning-based personality assessment tools' compliance with the practical and legal standards for use in practice (such as inter-algorithm reliability, test-retest reliability, and differential prediction across demographic groups). Additionally, we emphasize that the goal of machine learning-based personality assessment tools should not be to simply maximize the prediction of personality ratings. Rather, we should explore ways to use this new technology to further develop our fundamental understanding of human personality and to contribute to the development of personality theory.

Key words : personality measurement, machine learning, social media text mining, validity.

† 제1저자: 옥지수, 부산대학교 경영학과, jisoo.ock@pusan.ac.kr
‡ 교신저자: 안혜련, 부산대학교 경영학과, gpfus0618@pusan.ac.kr

The development of organizing structures of personality has led to an effective classification of a wide array of personality traits into dimensions that are commonly understood. Namely, the Big Five (Goldberg, 1990) and the six-factor HEXACO (Lee & Ashton, 2004) models are widely accepted as fundamental organizing structures of personality when conducting and considering personality research in organizational settings. Both the Big Five and the HEXACO models provide useful frameworks by which personality traits are labeled, defined, and measured, which has enabled researchers to develop a coherent body of empirical support for validity of personality measures and their relationship with important work-related behaviors and outcomes (e.g., job performance, teamwork, counterproductive work behavior, turnover).

Personality is typically measured using a scale(s) that consists of a series of items that describe different types of behaviors or dispositional tendencies that are theoretically and empirically associated with the measured trait. Respondents are often asked to subjectively evaluate the degree to which the statement described in each item is an accurate reflection of themselves (for self-reported measures) or of a target person (for observer-reported measures) on a Likert-type scale. Then, respondents' scores on the measure are derived through a linear combination of scores for each item (e.g., unit-weighted composite of item scores) or by estimating the latent commonality among the indicators through confirmatory factor analysis.

Although there is a broad consensus that personality meaningfully predicts important behaviors and outcomes in organizational settings (Barrick, 2005), the support for the use of personality tests in personnel selection settings is not unequivocal (e.g., Morgeson et al., 2007). Namely, both critics and proponents of personality testing have voiced concerns that self- or observer-reported ratings of personality can be susceptible to different types of biases (e.g., erroneous self-perception, friendship bias) and intentional response distortions (e.g., respondents presenting themselves in a socially desirable manner) that can undermine the validity and the practical usability of personality tests, especially in high-stakes personnel selection settings where job applicants have a clear motivation to present themselves in a positive manner. The self- and observer-report measures of personality are generally more efficient than relying on more objective and behaviorally-oriented approaches to personality assessment (which have their own issues regarding validity and reliability), but there also has been occasional doubts about whether people's subjective evaluations about themselves (or others) can be considered an appropriate standard for measuring personality (Boyd & Pennebaker, 2017), especially in applied settings for making high-stakes decisions. Perhaps as a result, organizations are quickly rushing to adapt new methods of personality assessment

that take advantage of the richness of big data and the analytical power of machine learning that are touted as being allegedly free from human errors and biases.

As test vendors develop and market their own versions of various machine learning-based psychological assessment tools, there is an increasing need for researchers to provide empirical evidence that inform their usability (although ideally, the order would be in reverse). In fact, there is a prevalent concern among measurement researchers that organizations might be overlooking the critical need for empirical evidence that support the psychometric soundness, construct validity, fairness, and legal defensibility of machine learning-based psychological assessment tools that inform their use in practice. Thus, we believe it is timely for a comprehensive review paper to summarize the current state of research and to identify the existing gaps in machine learning-based personality assessments that future research should explore.

The current paper provides a brief review of the current research on social media text mining and its application to personality assessment. Specifically, we integrate major discussions in recent reviews that emphasize the need for psychometric and theoretical validity evidence of big data personality assessment methods (Alexander et al., 2020; Bleidorn & Hopwood, 2019; Tay et al., 2020) with technical issues that researchers and practitioners need to consider in conducting text mining research (e.g., social media text analysis methods, text preprocessing). Additionally, throughout the paper, we provide readers with references to various user-friendly softwares and guidelines that readers can consult in conducting their own text mining research. Finally, based on our review, we offer some suggestions about future research directions in machine learning-based personality assessment that can inform both theory and practice.

Method for Literature Review

We searched prominent journals in applied psychology, psychometrics, personality, and research methods (e.g., *Journal of Applied Psychology, Journal of Personality and Social Psychology, Organizational Research Methods, Psychological Assessment, Psychological Methods*) for relevant references to include in our review using multiple combinations of keywords like "social media text mining," "machine learning," "natural language processing," and "personality assessment." These search terms returned 127 articles. We nominated 25 articles (30%) for consideration to include in the review based on our judgment of relevance to the topics examined in the review by reading the title and abstracts. We also examined recent reviews on the topic of machine learning approach to personality assessment (e.g., Tay et al., 2020) and text mining (Kern et al., 2016) for relevant

Table 1. Summary of Issues Reviewed and Major References

Topic	Description	Reference
Psychometric and construct validity	Summary of current research and call for more evidence that support the practical and legal viability of machine learning-based assessment tools for use in practice	Alexander et al. (2020) Bleidorn & Hopwood (2019) Stachl et al. (2020) Tay et al. (2020)
Theory advancement	Using machine learning to advance knowledge about human behavior and personality theory	Alexander et al. (2020) Tay et al. (2020)
Approaches to text analysis	Closed- vs. Open-vocabulary approach to text analysis	Eichstaedt et al. (2020) Kern et al. (2016)
Text preprocessing	Text transformations and their implications for text mining analysis results	Banks et al. (2018) Hickman et al. (in press)

Note. Full reference for the cited works are available in the references section.

materials and articles to review. Table 1 provides a summary of the topics that are reviewed and useful references that we examined with respect to each topic.

Machine Learning Approach to Personality Assessment

As we enter the digital age, the advancement of technology, computational power, and statistical and quantitative techniques are giving new ways for researchers to collect, assess, store, share, and analyze large and complex human behavioral data that used to be difficult to access and explore (Woo et al., 2020). These advancements are having an important and immediate impact on personality assessment and its application. Namely, there has been a major influx of research and interest on personality scale development that apply machine learning on digital records from a wide range of sources, including social media behaviors (e.g., Kosinski et al., 2013; Youyou et al., 2015), social media languages samples (e.g., Park et al., 2015; Schwartz et al., 2013), financial transactions (e.g., Gladstone et al., 2019), personal weblogs (e.g., Iacobelli et al., 2011), and smartphone data (e.g., Chittaranjan et al., 2013).

Machine learning-based personality assessment typically involves gathering a large amount of digital behavior records, which are then used to create indicators or scales that maximize the prediction of individual differences in personality as measured by traditional self- or observer-reported personality measures (Bleidorn & Hopwood, 2019; Stachl et al., 2020). Although research in this area is still at a relatively early stage, results have shown that computer-based

assessments of digital behavior are useful predictors of personality. For example, Youyou et al. (2015) made computer-based judgments of personality using patterns of Facebook "Likes" for over 70,000 participants and found that they accurately predicted self-ratings of Big Five personality factors measured using the 100-item International Personality Item Pool (IPIP; Goldberg et al., 2006) questionnaire. The results even showed that the computer-based personality assessments were more strongly correlated with self-rated personality scores (r = .56) than the average personality ratings obtained from friends of the participants (r = .49). In other words, computer-based judgments were at least as accurate (if not slightly more accurate) with human-based judgments in predicting personality.

The guiding principle for using digital records of behavior as indicators of personality is that people's behavior in online environments are reflective of their attitudes, preferences, interests, and tendencies that are largely consistent with their personality (Back et al., 2010). For example, it has been shown that extraverted individuals tend to have more friends on social networking sites, update their online profiles more often, and have deeper social networks (Kosinski et al., 2014). These digital behavior patterns align with extraverted individuals' tendency to be more social, outgoing, and gregarious. In many ways, because online activities are generally self-manifested and less likely to be affected by impression management, they can potentially provide less biased information about personalities than traditional measurement tools that rely on subjective evaluation of the self or of an observer (Kosinski et al., 2014). In addition to the innovativeness of such novel approaches to personality assessment, it is suggested that these new tools allow for the assessment of psychological constructs in an unobtrusive and bias-free manner that allegedly offer improved validity and fairness beyond those provided by traditional methods of assessment. However, research has lagged behind in the provision of empirical support for such claims (we will discuss this issue in more detail).

Among the different sources of digital records, social media platforms offer big data that are particularly useful for personality researchers (Stachl et al., 2020). First, the demographics of the people who use social media are highly diverse in terms of race, gender, nationality, culture, and so forth. Second, most people who actively use social media tend to do so on a regular basis (Lenhart et al., 2015; Perrin & Anderson, 2019). As a result, social media records provide data that are not only large and representative of demographics at the level of populations (which is more difficult to achieve using smaller samples typically collected for research studies), but they also tend to provide intensive longitudinal samples of online behavior (often multiple time points per day) that are more generalizable and shed brighter light on

the malleability of personality over time. As a result, digital records from social media platforms, and social media text data in particular, have been popularly used in the big data personality research (Alexander et al., 2020; Tay et al., 2020).

Social Media Text Analysis Methods

Broadly, there are two approaches to conducting social media text analysis: 1) closed-vocabulary approach; and 2) open-vocabulary approach (Kern et al., 2016). The main distinction between these two analytic approaches is in the degree to which the text analysis process is automatized.

In closed-vocabulary text analysis, text data are assigned into psychosocially relevant categories of words that are pre-determined based on theory to reflect certain emotions or sentiments (Eichstaedt et al., 2020). For example, words like *sad*, *anger*, and *hate* may be theorized as being part of a *negative emotions* category because they commonly reflect various negative emotional states. These pre-determined word categories, which are called "dictionaries," are incorporated into computer, which then scans the digital text information, categorize each word into different dictionaries, count the number of times the words from each dictionary has occurred, and calculate their frequencies as outputs to be used in subsequent statistical analyses (Eichstaedt et al., 2020).

One of the most commonly used closed-vocabulary text analysis software is the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015). The default dictionary in LIWC contains over 6,400 commonly used English words, word stems, and even some emoticons that are classified into pre-determined dimensions according to psychological constructs (e.g., anger, sadness) and parts of speech that they represent (Seih et al., 2020). Users also have the option to define and add their own dictionaries into the analysis. When users import text files they want to analyze onto LIWC, it classifies the words in the documents into the pre-determined word categories and provides outputs in terms of the percentage of total words that are classified into different word categories (Seih et al., 2020).

In addition to LIWC, an increasing number of dictionary-based text analysis options are becoming available in open-source statistical environments like R (e.g., syuzhet package; Jockers, 2020). The challenge in taking advantage of such tools is overcoming the technical barriers associated with using such programming languages, especially for researchers who are less familiar with programming. To overcome this challenge, experts are introducing useful guidelines (e.g., Welbers et al., 2017) and new user-friendly applications (e.g., topicApp; Banks et al., 2018) that allow users easy access to text analysis techniques even without extensive experience in programming language.

In open-vocabulary text analysis, linguistic

features of the text (e.g., words, phrases, topics discussed) are automatically analyzed to formulate word clusters or *n*-grams (word length in a phrase) that are semantically related within a large text data (Eichstaedt et al., 2020). These clusters can then be used as predictors to find word features that show strongest convergence to a trait, behavior, or other outcomes of interest. Unlike closed-vocabulary text analysis, which is dependent on theoretically determined word categories based on linguistic, psychological, and social theories, open-vocabulary text analysis is largely data-driven. Because open-vocabulary text analysis is more flexible in terms of determining semantic word clusters from a given text data, it tends to be more effective compared to closed-vocabulary text analysis in accurately interpreting language information that are more nuanced, subtle, and ambiguous in their meaning (Eichstaedt et al., 2020).

For example, the differential language analysis (DLA) approach is a frequently used open-vocabulary text analysis method in identifying psychological characteristics that underlie text data (Schwartz et al., 2013). DLA approach follows a regression framework where the relative frequency of language features (e.g., word, phrase, topic) that are identified as indicators is used to predict the dependent variable of interest (e.g., personality trait; Kern et al., 2016). The number of regression analyses that are conducted for each indicator easily reaches thousands or more depending on the complexity of the input text data.

Because of the sheer volume of data that is typically processed in open-vocabulary text mining, DLA results in thousands of correlations between indicators and outcomes that are small in terms of effect size (r = .00 to .20), even after removing indicators that are used only in a very small subset of the sample (Kern et al., 2016). Thus, it is recommended that conservative p-value adjustments are made (e.g., Bonferroni correction) in order to avoid over-interpretation of statistically significant results, especially given the large sample size that is typically involved in machine learning.

Validity Evidence for Personality Assessment Using Social Media Text Mining

Majority of research on social media text mining for personality assessment has focused on examining the convergent validity of scales derived from social media text mining on self- (or observer-) reported measures of personality (mostly for measures of the Big Five). Results generally support that social media text mining algorithms can be used to predict personality, but the text mining method employed adds a meaningful variability in the strength of the prediction, especially for open-vocabulary approach.

Namely, Tay and colleagues (Tay et al., 2020) conducted a meta-analysis of convergent

validity evidence and found that both closed- and open-vocabulary approaches showed similar levels of convergent validity (average correlations ranging between r = .21 and .29 across Big Five traits). The meta-analysis included only a single study for closed-vocabulary text analysis approach. However, this was a very large scale study (Schwartz et al., 2013) that analyzed more than 700 million words, phrases, and topics from Facebook messages of a very large sample (sample size around 75,000). They found that word categories in closed-vocabulary approach showed significant correlations with self-reported measurement scores on the Big Five personality traits (correlations ranging between r = .21 and r = .29; see Table 2).

Similar levels of sample weighted mean correlations were found for open-vocabulary text analysis across the Big Five personality traits (mean r = .27 and ranging between r = .24 and r = .29; see Table 2). However, there was an important difference in the magnitude of convergence validity across open-vocabulary text analysis methods. Namely, the sample-weighted convergent validity estimates were made primarily from samples in Golbeck (2016) and Park et al. (2015). In Golbeck (2016), a commercial product called *Receptiviti* that combines the features of both closed- and open-vocabulary text analysis approaches was used to estimate respondents' Big Five personality from social media language use,

Table 2. Convergent validity of computer-based personality assessment from social text mining with self-reports of personality (from Tay, Woo, Hickman, & Saef, 2020)

Assessment approach	Personality trait	*k*	*N*	*r*	SD_r	*95% CI*
Closed vocabulary	Extraversion	1	18,177	.27		
	Agreeableness	1	18,193	.25		
	Conscientiousness	1	18,195	.29		
	Emotional stability	1	18,177	.21		
	Openness	1	18,202	.29		
Open vocabulary	Extraversion	5	13,893	.29	.11	[.15, .43]
	Agreeableness	4	13,589	.28	.08	[.15, .41]
	Conscientiousness	4	13,589	.26	.10	[.10, .42]
	Emotional stability	5	13,893	.24	.09	[.11, .36]
	Openness	4	13,589	.28	.14	[.06, .50]

Note. k = number of studies included in the meta-analysis; N = total sample size; r = sample-weighted mean correlation between computer-based personality assessment and self-reports; SD_r = standard deviation of correlation estimate; 95% CI = 95% confidence interval.

whereas Park et al. (2015) developed a text mining approach for personality prediction from a very large social media text data.

More specifically, Park et al. (2015) collected Facebook status messages (brief text messages that users post on their profiles) from a very large sample of Facebook users (N = 71,556) who allowed their status messages to be accessed. Park et al. (2015) captured messages that these users posted in a time frame between January 2009 and November 2011, resulting in over 15 million messages in total. Participants also completed measures of the Big Five using IPIP items (Goldberg et al., 2006). Applying open-vocabulary text analysis method (latent Dirichlet allocation; LDA) to a large model development sample (n = 66,732), Park et al. (2015) extracted a large set of predictors (text features predictive of self-reported personality; initial set of p = 51,060) that allowed the model to detect even the subtle (but informative) signals that are more difficult to extract in smaller text data (Bleidorn & Hopwood, 2019).

The convergent validities found in Park et al. (2015) were high and comparable to values typically found in other monotrait-multimethod correlations: r = .43 for Openness, r = .37 for Conscientiousness, r = .42 for Extraversion, r = .35 for Agreeableness, and r = .35 for Emotional Stability. These correlations were higher than the correlations in Golbeck (2016) and other studies that employed open-vocabulary text analysis approach for personality assessment (as reported in Table 2). Although research efforts for empirically examining the convergent validity of social media text mining approaches for predicting personality traits is still at a very early stage, these results suggest that personality prediction from social media text data is certainly possible.

However, much more information about the psychometric properties and validity of such computer-based personality assessment beyond convergent validity with traditional self- or observer-reported personality test scores is needed to further refine and enhance the utility of machine learning-based personality assessment scales (Bleidorn & Hopwood, 2019). Although such evidence remains scant, Park et al. (2015) is a notable exception.

Namely, in addition to convergent validity, Park and colleagues examined the test-retest reliability of personality assessment data from Facebook users' online posts by correlating within-person predictions of Big Five personality from digital text information in different time points. Specifically, Park et al. (2015) split the participants' Facebook posts into four consecutive 6-month intervals, retaining data for those who had written at least 1,000 words within each pair of intervals (e.g., users who wrote at least 1,000 words on their Facebook posts during Interval 1 and Interval 2). For each 6-month interval, they calculated personality predictions based on text data within each interval and

then calculated the correlations between the predictions for each trait across all possible pair of intervals (e.g., correlating Conscientiousness prediction at Interval 1 with Conscientiousness prediction at Interval 2). This resulted in six test-retest correlations for each trait (correlations between predictions for Intervals 1 through 4). The average test-retest correlations across the five traits across all combinations of intervals was r = .70, which is comparable to the test-retest reliability values typically found for self-reported personality measures over similar periods of time.

The test-retest reliability evidence for personality assessment from social media text mining indicates that there are meaningfully reliable linguistic data available on social media websites that allow meaningful inferences to be made about users' personality. Namely, because people may use social media somewhat casually without too much deliberation about what to express on these platforms, one could argue that much of the contents could reflect random, spur-of-the-moment thoughts that might be transient and inconsistent over time. However, the high levels of test-retest reliability estimates across the Big Five traits found in Park et al.'s (2015) study provides evidence that meaningful proportion of social media text can be used to infer stable components of one's personality.

In addition to test-retest reliability, Park et al. (2015) examined the discriminant validity of Big Five measures derived from social media text mining. Namely, they examined the intercorrelations among the Big Five traits from digital assessments and found that the mean correlations were significantly higher compared to the intercorrelations typically found in self-report measures (average r = .29 vs. r = .19, respectively), suggesting relatively lower discriminant validity evidence for social media text mining approach.

The lower discriminant validity evidence found in Park et al. (2015) reflects one of the unique challenges of machine learning-based personality assessment. Namely, to maximize convergence, researchers often include all informative digital text indicators into the analysis. As a result, the same text indicators may be used to predict different personality traits, resulting in lower discriminant validity problem (Bleidorn & Hopwood, 2019). Relatively simple ways to alleviate this issue would be to use different corpus for predicting different personality traits or to focus on a smaller number of construct-relevant text information that are determined *a priori* in running the prediction model (much like in closed-vocabulary approach). As Campbell and Fiske (1959) observed in the earlier days of psychological assessment, there is a need to move beyond the current focus on examining test score convergence to also identifying evidence of discriminant validity in establishing construct validity for personality assessment using social media text mining.

Finally, Park et al. (2015) examined the criterion-related validity of social media text

mining personality assessments for predicting various external criteria (e.g., life satisfaction, physical symptoms) that were self-reported. In general, criterion-related validity values for social media text mining personality assessments were lower than those for self-reported measures of the Big Five. However, even without the common method variance that is inevitably present in criterion-related validities for self-reported measures, personality assessments from social media text mining showed meaningful correlations with self-reported measures of external criteria. For example, social media text-based measures of Agreeableness, Conscientiousness, and Neuroticism showed modest correlations with life satisfaction (r = .21, .19, and -.19, respectively).

The validity evidence in Park et al. (2015) provide further support for personality prediction from social media text data. However, as previously mentioned, more research effort is needed to further refine and improve the utility of machine learning approach to personality assessment, particularly with respect to their application in applied settings (we discuss this issue in more detail later).

Text Preprocessing for Conducting Text Analysis Research

In addition, the practical implications of different decisions that could be made during text preprocessing need to be explored further. Text preprocessing refers to the general process of making corrections or cleaning of the text data so as to enhance measurement precision and predictive accuracy of text analysis process (Banks et al., 2018; Hickman et al., in press). Just as researchers engage in data preprocessing before delving into statistical analyses of data (e.g., outlier analysis, identifying careless responses), preprocessing is a standard practice in text mining. This section focuses on some of the preprocessing methods that can critically affect the accuracy of text analysis and important issues to take into consideration in undertaking text preprocessing.

There are several different types of text preprocessing methods that researchers need to consider in conducting text analysis. Namely, in English text analysis, lowercase conversion refers to converting all letters in the text to lowercase. In computer language, uppercase and lowercase words are treated as being distinct. So even though there is no semantic difference between a word that is capitalized and the same word that is not (e.g., “Personality” and “personality”), they will be stored as distinct words in text mining. This adds unnecessary complications to data dimensionality and leads to decreased statistical power. As a result, it is generally recommended that lowercase conversion be always used in text mining (Banks et al., 2018; Kobayashi et al., 2018).

Studies have consistently shown that lowercase conversion leads to improved prediction accuracy

(Kobayashi et al., 2018). For example, lowercase conversion has been shown to improve the rate at which computers are able to accurately identify spam e-mail messages and classify news stories into appropriate categories (Uysal & Gunal, 2014). In fact, most research in organizational science domains that have used open-vocabulary text accuracy reported using lowercase conversion as part of their preprocessing, and lowercase conversion is the default option in many text mining softwares and open source software packages that conduct text mining (Banks et al., 2018; Hickman et al., in press). However, because uppercase vs. lowercase letters are often not distinguished in text mining, uppercase words that are important might be erroneously identified as irrelevant if they are not properly identified. For example, if the document refers to the field of information technology as "IT", each instance of "IT" would be identified as the word "it" and likely be removed from analysis. To avoid such errors, key words that are spelled the same way with commonly occurring irrelevant words should be identified and made distinguishable prior to the analysis.

Another important and common preprocessing technique is the handling of negation. When words are preceeded by negation (e.g., *no*, *not*, *never*), it usually alters the meaning of the word or the phrase. For example, the semantics of a phrase "*not happy*" is clearly different than when the words "*not*" and "*happy*" are separately treated as independent words and interpreted independently. Yet, without proper handling of negation, each instance of "*not happy*" will be counted the same as an instance of "*not*" and an instance of "happy" and fail to record the proper meaning of the text (Hickman et al., in press). A simple (but effective) way to address this issue is to append a special character to each negation so that text analysis will distinguish between a word with vs. without negation (e.g., replacing "*not*" with "*not_*", so that instances of "*not happy*" will be replaced with "*not_happy*"). A review of text preprocessing techniques has shown that handling of negation improved the accuracy at which semantic interpretations were made through text mining (Hickman et al., in press). For example, Smith et al. (2015) found that handling negation increased the level of accuracy in which computers handled question-answering tasks based on fictional stories.

Relatedly, other forms of text transformation might be needed to accurately account for semantic information in the text. In many languages, words can have multiple meanings or uses and the meanings of words may be contextually dependent. Namely, in the Korean language, there are several instances where a word can take on a very different meaning depending on how it is used or what words precede or follow it. For example, the word "난다" could mean "to fly" when it is used alone, but it could also mean (roughly translated) "to

display" or "to express an emotion" when it is used in a different context (e.g., "눈물이 난다", "화가 난다"). Machine learning can effectively distinguish between such words if they frequently co-occur with a different word. In other words, if the word "눈물이" frequently co-occurs with "난다", computers can distinguish between instances of "눈물이 난다" with other instances of "난다" that appear within the corpus. Also, it may be useful to set a minimum frequency of the number of times a word or a phrase has to appear within a corpus to be included in the analysis. Doing so can minimize adding unnecessary complexity to text analysis, such as distinguishing the exact semantic meaning of a word that only sparsely occurs and is not a meaningful part of the text. Otherwise, researchers can manually code for such differences in meanings of the words (e.g., creating a custom dictionary). On the whole, in text mining, researchers need to take great care in text preprocessing to account for these types of subtle differences in the text that could vastly change their semantics and analytic results.

There are other types of text preprocessing steps that intuitively make sense to implement. For example, correcting spelling errors are needed to improve the accuracy and interpretability of the text. However, increased accuracy and interpretability of the text may not always be a good thing in text analysis. For instance, if the goal of the analysis is to make inferences about individual differences in personality using text data, making corrections to spelling errors might lead to a loss of important information about individual differences in linguistic behavior that may be relevant to personality. For example, people who are more likely to make spelling errors might do so because they are less attentive or less organized, which may reflect lower Conscientiousness. Thus, although preprocessing methods might be effective for improving text interpretation and ensuing predictions that are made based on text analysis, the decision to implement each type of preprocessing should be preceeded by a careful consideration of the purpose of the analysis and whether the implications of the preprocessing procedure can help (or undermine) the goal of the analysis.

In addition to the discussion presented in the current paper, we refer readers who are interested in conducting organizational text mining research to Hickman et al. (in press) for an in-depth review of the issues in text preprocessing for organizational text mining research and recommendations for their appropriate use and reporting standards. Also, we refer readers to a useful R package (preText; Denny & Spirling, 2018) and Shiny application (topicApp; Banks et al., 2018) that allow users to run a diagnostics test of the effect of different text preprocessing decisions on the inferences that are drawn from the analysis.

Questions for Future Research

Concomitant with the increasing prevalence of machine learning technology and its application to personality assessment, there has been an accumulation of empirical evidence that support the validity of machine learning-based personality assessments. However, there is still a number of questions that need to be explored further to improve the applied use of social media text mining for personality assessments and to inform personality theory.

Future Research for Improving Application

Namely, there is an important gap in the literature with regards to examining the criterion-related validity of social media text mining personality assessments for predicting job performance. A key determinant of the usefulness of a selection measure is whether it can be used to predict effective job performance (or other behaviors or outcomes of interest to the organization). Although Park et al. (2015) demonstrated that social media text mining personality measures can be used to predict external variables, those variables were largely irrelevant to the types of criterion that would be considered in personnel selection contexts. Thus, to facilitate the use of text mining-based personality assessment tools in personnel selection, there is an important and immediate need for empirical evidence that demonstrates that those measures can be used to predict effective job performance.

Also, with regards to the reliability of machine-learning based personality assessments, the effect of rater (or in the case of machine learning, algorithm) is an important source of error that would be useful to identify and model (Sajjadiani et al., 2019). As mentioned, different text analysis methods apply different algorithms for deriving semantically related word clusters within text data, which could produce meaningfully different predictions across different methods even when they are applied to the same text data. The difference in prediction across multiple methods is akin to the difference in evaluation across multiple raters that is modeled as error variance in interrater reliability, and is relevant and important to identify in social media text mining approach to personality assessment. Thus, future studies should examine the “inter-algorithm” reliability of different text analysis approaches for personality assessment, especially given the likelihood that researchers without expertise in machine learning (yet, could greatly gain from the application of machine learning) are less likely to be familiar with the specific iterative processes that are involved in different text analysis methods.

In addition, because of the data-driven nature of open-vocabulary text analysis, cross-validation of results are essential to avoid overfitting a model that is developed on a specific sample. In big data analysis, prediction model is developed on a subset (or multiple subsets) of data (called

training data), and the performance of the developed model is tested and further refined based on how the model performs in an independent subset (or multiple subsets) of data (called test data). That is, prediction model is developed based on a subset(s) of data, then cross-validated on other subset(s) of the data to form a final model that is generalizable. However, a less frequently explored question is the degree to which a prediction model that is developed on a specific corpus is cross-validated on another corpus. For example, does a model that is developed based on one digital text big data generalize to another digital text big data that is collected from a different sample and/or time? What about across different platforms (e.g., personality prediction from models built using digital text information in Twitter vs. Facebook vs. LinkedIn) or document types (e.g., social media text mining vs. resume text mining)? More research that explores such questions are needed because the data-driven nature of the derivation of predictors from big data make it difficult to understand *why* prediction occurs. Without this understanding, there is a need to remain cautious about the robustness and the implications of model prediction results that are made based on digital text mining.

In addition to understanding the validity and psychometric soundness, for social media text mining personality assessments to be used in applied settings, there is an important need for research demonstrating that they (and machine learning-based assessments in general) comply with the legal standards for use in personnel selection. Professional test standards (AERA, APA, & NCME, 2014; SIOP, 2018) specify that in addition to psychometric properties (e.g., reliability, construct validity), measures that are used in personnel selection should demonstrate evidence that informs demographic group mean differences (which has implications for adverse impact) and fairness (e.g., equivalent reliability and criterion-reliability across groups). The goal of these recommendations is to help organizations both practically, in improving the quality of personnel selection decisions, and legally, in preventing any potential violation of laws against unfair discrimination in the hiring process. In the absence of sufficient supportive evidence for such recommended measurement properties, the application of measures in selection process can have serious negative practical and legal implications for organizations.

Because so much remains unknown about machine learning-based assessments in general (what they measure, how the extracted indicators are scored, and so forth) and because of the open-ended nature of algorithms in machine learning, machine learning-based assessments are especially prone to violation of legal requirements for selection. Namely, it is possible that algorithms may extract and use legally inappropriate data (e.g., race, gender, age) in scoring job applicants (Tonidandel et al., 2018).

For example, Amazon abandoned its use of their artificial intelligence selection tool for hiring engineers after learning that it was trained to reflect the male-dominant nature of the technology industry, which then penalized female engineers based on resume information that suggest that the applicant is female (Dastin, 2018). Such discriminatory scoring algorithms can easily develop without proper supervision, especially given that job performance ratings, which algorithms are trained to predict in machine learning, are also known to contain various types of biases, fallacies, and errors due to performance-irrelevant external factors (Murphy, 2020). The greater the degree to which criterion ratings that contain such fallacies (especially those that unfairly discriminate against certain groups of individuals) are regarded as the "ground truth" in terms of job performance, algorithms that are trained to maximize the prediction of job performance ratings are more likely to make discriminatory decisions that increase the likelihood of adverse impact (and the possibility of litigation if adverse impact is found) and potentially deny employment opportunities to qualified individuals.

In sum, it is essential for organizations to demonstrate evidence that selection test batteries that are used in selection minimize the potential for adverse impact and that the use of those tests are justified through job analytic and construct validity evidence that support their use in the employment process (Gonzalez et al., 2018). However, the current state of research is significantly lagging behind the popularity of machine learning-based assessments in practice. As a result, organizations are having to rely on claims from test vendors that vary in terms of empirical support for those claims or small-scale evidence that may be sample- and/or organization-specific (Oswald et al., 2020). When organizations adapt psychological tests with such limited support for their use, it poses substantial threat to the organization that use these measures (by increasing the likelihood of adverse impact that undermines the legal defensibility of selection) and to our society at large (by systematizing the patterns of bias and discrimination against stigmatized group members). To stand up to these challenges and to add confidence to the use of modern methods of psychological assessments that are quickly becoming widely prevalent, there is an immediate need for research that address these issues before these modern methods of assessments are further disseminated in practice.

Finally, researchers should take advantage of these new methods of assessment to examine new and unexplored psychological factors that might underlie effective job performance. In the wake of dynamic changes that are taking place in the business world and the wide range of threats and opportunities they pose to organizations, it is becoming increasingly important for organizations to be able to develop and maintain a skilled and adaptable

workforce. In this environment, there might be psychological factors that are central to successful job performance that are not necessarily traditionally considered in personnel selection, or they are considered but not reliably measured.

Of course, the process of identifying such psychological factors should not be purely data-driven. Rather, there should be efforts to develop a coherent theory about what those psychological factors are, how they might be reflected in various unstructured data, assess the degree to which algorithms can be developed to effectively measure them using such data, and examine whether those measures are in fact predictive of effective job performance. As exhaustive and complicated such process can be, it could also provide organizations with valuable pieces of evidence that they can use to further improve the quality and the utility of personnel selection decisions.

Future Research for Improving Theory

In addition, there are outstanding questions that are more fundamental to the use of machine learning-based personality assessment and its application that need to be carefully considered. First, there needs to be more in-depth theoretical discussions regarding which aspects of personality are being measured in social media text mining. Personality scores that are generated from social media text mining are fundamentally different than personality scores in traditional self- or peer-reported personality measures. Namely, the Trait-Reputation-Identity Model (McAbee & Connelly, 2016) proposes that personality variance is composed of unique perceptions that individuals have about themselves (identity), impressions that individuals convey to the public that are agreed upon by other people (reputation), and commonality between identity and reputation that reflect consensus about underlying traits (trait). In this framework, self-reported personality ratings can be said to reflect an individual's identity and observer-reported personality ratings can be said to reflect an individual's reputation. However, it is not very clear which aspect(s) of personality variance is reflected in personality measurement scores that are derived from social media text mining (Tay et al., 2020).

Some features of social media behaviors, like online posts about one's feelings and emotions, and expressions of endorsement or opposition about certain news events or products, can be argued to reflect introspective accounts of people's own identity. It could also be argued that the contents of external online language behaviors (e.g., online conversation with others via comments) are driven by social dynamics that prompt others to interact with the individual in a certain way according to the perceived image that others have about him/her. However, such distinction becomes difficult to assess in text mining because algorithms largely determine which features of text information are extracted and how they are scored (and often

unbeknownst to the researchers conducting the analysis, which is often referred to as the "black box" problem in machine learning), with the ultimate goal of maximizing their correlations with self- and/or observer-reported personality scores. In this context, concerns about the specific aspect(s) of personality variance being captured becomes somewhat irrelevant, so long as scoring algorithms can be trained to maximize prediction of the outcome. As a result, although personality assessments from social media text mining can be used to predict personality scores from traditional measurement approaches, it is often difficult (if not impossible) to clearly understand what the computer-based scores actually reflect about the measured individuals.

Similarly, even when meaningful predictions about people's personality are made based on social media text data, it is often difficult to identify why those predictions are made. For example, Park et al. (2015) found that *n*-grams on Facebook posts that showed strongest prediction for Extraversion contained pronouns (e.g., I can't, it is), prepositions, (e.g., from, of, into), and articles (e.g., the, as) (Hickman et al., in press). However, it is not immediately clear why extraverted individuals would be more inclined to use these words or word phrases in their social media posts compared to individuals who are less extraverted.

Although we have begun to use machine learning as a useful tool for predicting personality, we have not yet been able to use machine learning to gain deeper insights into the complexities of human personality, largely because we have not been able to understand what features of text data are consistently used to predict personality in text mining, and to the extent that we have been able to find reliable prediction, we have not been able to understand why such predictions are made. Obviously, these limitations need to be addressed. Namely, the question needs to move beyond *whether* we can make inferences about people's personality from social media text mining to understanding *why*, and based on this understanding, move to *how* we can improve them.

Second, it is important to understand that traditional self- or observer-reported personality measurements are imperfect indicators of personality themselves. Thus, even if machine learning-based personality assessments provide reasonable prediction of self- and/or observer-reported personality, they too inevitably contain various types of systematic (e.g., self-enhancement bias in self-reported ratings, contamination due to social relations in observer-reported ratings) and random measurement error variances that make them imperfect representations of personality true scores. Thus, more accurate prediction of traditional personality measurement scores from machine learning-based personality assessment may not necessarily mean more accurate assessment of the personality trait(s) that the assessment scale intends to measure.

In that regard, future research on machine learning-based personality assessment should move beyond simply focusing on maximizing prediction of personality scores as measured by traditional personality assessment methods. Rather, the focus should be demonstrating whether, and to what extent, machine learning-based personality assessments are able to provide incremental information about personality above and beyond traditional measurement approaches. With respect to their application to practical contexts, such as personnel selection, we need to examine the degree to which machine learning-based personality assessments can offer incremental validity over traditional measurement methods in predicting job performance, and more importantly, understand why incremental prediction (if any) occurs.

Part of the reason for the interest and excitement towards machine learning approach to personality assessment is for this very expectation that it can be used to capture more subtle and even concealed aspects of human personality (e.g., negative personality traits) in a manner that is less prone to human errors, biases, and cognitive fallacies, which should lead to improved prediction. To continue to improve the research on machine learning-based personality assessment, we as a field need to engage in more in-depth theoretical discussions about what should be considered the "ground truth" about one's personality, how it can be captured, and predicted using machine learning approach.

Concluding Comments

With the advent of machine learning-based personality assessment tools, we are seeing some of the same issues being raised that were also raised in the "good old daze" of profuse number of personality constructs and measurements (Hough, 1998). We need history to repeat itself. The field of psychology has left the early days of dust-bowl empiricism when a measure was deemed useful so long as it predicted any outcomes of importance. Now, in the era of theory-driven research, we need to examine whether machine learning-based personality assessment scores can withstand the rigor of fundamental construct validation process. Moreover, research needs to quickly catch up to practice by examining whether the applied use of machine learning-based personality assessments (e.g., high stakes personnel selection) provide results that are effective, fair, and legally appropriate. In that regard, the past and present development in personality theory and measurement can serve as a useful guiding principle for the direction of the future of machine learning-based personality assessment and its application to practice.

The current paper brings together the latest research from multiple areas of social media text mining approach to personality assessment. We hope that readers will find the current paper helpful in developing an integrative understanding and appreciation for various issues

that should be taken into consideration in both research and applied aspects of text mining approach to personality assessment (and machine learning approach to psychological assessment in general).

References

Alexander, L., III, Mulfinger, E., & Oswald, F. L. (2020). Using big data and machine learning in personality measurement: Opportunities and challenges. *European Journal of Personality, 34,* 632-648. https://doi.org/10.1002/per.2305

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science, 21,* 372-374. https://doi.org/10.1177/0956797609360756

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology, 33,* 445-459. https://doi.org/10.1007/s10869-017-9528-3

Barrick, M. R. (2005). Yes, personality matters: Moving on to more important matters. *Human Performance, 18,* 359-372. https://doi.org/10.1207/s15327043hup1804_3

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review, 23,* 190-203. https://dx.doi.org/10.1177/1088868318772990

Boyd, R. L., Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences, 18,* 63-68. https://doi.org/10.1016/j.cobeha.2017.07.017

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105. https://doi.org/10.1037/h0046016

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing, 17,* 433-450. https://doi.org/10.1007/s00779-011-0490-1

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis, 26,* 168-189. https://doi.org/10.1017/pan.2017.44

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan,

C. A., Tobolsky, V., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2020). Closed and open vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. https://doi.org/10.31234/osf.io/t52c6

Gladstone, J. J., Matz, S., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological Science, 30,* 1087-1096.https://dx.doi.org/10.1177/0956797619849435

Golbeck, J. A. (2016). Predicting personality from social media text. *AIS Transactions on Replication Research, 2,* 1-10. https://doi.org/10.17705/1atrr.00009

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216-1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84-96. https://doi.org/10.1016/j.jrp.2005.08.007

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). Where's the I-O? Artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions, 5,* 33-44. https://doi.org/10.25035/pad.2019.03.005

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (in press). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*. https://doi.org/10.1177/1094428120971683

Hough, L. M. (1998). The millenium for personality psychology: New horizons or good old daze. *Applied Psychology, 47,* 233-261. https://doi.org/10.1111/j.1464-0597.1998.tb00023.x

Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). Large scale personality classification of bloggers. In S. D'Mello, A. Graesser, B. Schuller, & J. Martin (Eds.), *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction* (pp. 568-577). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-3-642-24571-8_71

Jockers, M. (2020). syuzhet: Extracts sentiment and sentiment-derived plot arcs from text [Computer software manual]. https://cran.r-project.org/web/packages/syuzhet.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods, 21,* 507-525. https://dx.doi.org/10.1037/met0000091

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational Research Methods, 21,* 733-765. https://doi.org/10.1177/1094428117722619

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D.,

& Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning, 95,* 357-380. https://dx.doi.org/10.1007/s10994-013-5415-y

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research, 39,* 329-358. https://doi.org/10.1207/s15327906mbr3902_8

Lenhart, A., Duggan, M., Perrin, A., Steepler, R., Rainie, L., & Parker, K. (2015). *Teens, social media, & technology overview 2015: Smartphones facilitate shifts in communication landscape for teens* (p. 48). Retrieved from https://www.pewresearch.org/wp-content/uploads/sites/9/2015/04/PI_TeensandTech_Update2015_0409151.pdf

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review, 123,* 569-591. https://doi.org/10.1037/rev0000035

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K.., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Selection, 60,* 683-729. https://doi.org/10.1111/j.1744-6570.2007.00089.x

Murphy, K. R. (2020). Performance evaluation will not die, but it should. *Human Resource Management, 30,* 13-31. https://doi.org/10.1111/1748-8583.12259

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior, 7,* 505-533. https://doi.org/10.1146/annurev-orgpsych-032117-104553

Park, G., Schwartz, A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108,* 934-952. https://doi.org/10.1037/pspp0000020

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin. https://dx.doi.org/10.15781/T29G6Z

Perrin, A., & Anderson, M. (2019, April 10). *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018.* Retrieved May 27, 2019, from Pew Research Center website: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology, 104,* 1207-1225. https://doi.org/10.1037/apl0000405

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal,

M., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8,* e73791. https://dx.doi.org/10.1371/journal.pone.0073791

Seih, Y.-T., Lepicovsky, M., & Chang, Y.-Y. (2020). Your words reveal your thoughts: A two-wave study of assessing language dimensions in predicting employee turnover intention. *International Journal of Selection and Assessment, 28,* 484-497. https://doi.org/10.1111/ijsa.12302

Smith, E., Greco, N., Bosnjak, M., & Vlachos, A. (2015, September). A strong lexical matching method for the machine comprehension test. In *Proceedings of the 2015 Conference on the Empirical Methods in Natural Language Processing* (pp. 1693-1698).

Society for Industrial and Organizational Psychology (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). Bowling Green, OH: The Society for Industrial and Organizational Psychology.

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality, 34,* 613-631. https://doi.org/10.1002/per.2257

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality, 34,* 826-844. https://doi.org/10.1002/per.2290

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods, 21,* 525-547. https://doi.org/10.1177/1094428116677299

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50,* 104-112. https://doi.org/10.1016/j.ipm.2013.08.006

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures, 11,* 245-265. https://doi.org/10.1080/19312458.2017.1387238

Woo, S. E., Tay, L., Proctor, R. W. (2020). *Big data in psychological research*. Washington: American Psychological Association.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112,* 1036-1040. https://doi.org/10.1073/pnas.1418680112

투고일자 : 2021. 01. 11
수정일자 : 2021. 02. 12
게재확정 : 2021. 02. 27

한국심리학회지: 산업 및 조직
Korean Journal of Industrial and Organizational Psychology
2021. Vol. 34, No. 2, 213-236
https://doi.org/10.24230/kjiop.v34i2.213-236

머신러닝 기반 성격검사에 대한 문헌연구 및 향후 연구 방향에 대한 제안

옥 지 수 안 혜 련

부산대학교 경영학과

디지털 시대로 접어들면서 머신러닝을 통한 성격검사라는 새로운 유형의 성격검사방법이 관심을 끌고 있다. 머신러닝을 통한 성격검사는 사람들이 디지털 환경에서 보이는 실제 행동에 대한 데이터를 축적하고 여기에 알고리즘을 적용하여 개인의 성격을 예측한다. 이와 같이 머신러닝을 통한 검사는 사람의 판단이 아닌 데이터를 토대로 계산된 알고리즘을 통해 평가가 이루어지기 때문에 개인적 편견이나 인지적 오류에서 벗어나 정확하고 객관적인 평가를 가능하게 한다는 점이 강조된다. 이러한 이유 때문에 인공지능 면접과 같은 머신러닝 기반 평가도구들은 우리나라를 포함하여 전 세계적으로 빠르게 확산되고 있다. 하지만 이러한 평가도구들이 채용과정에 적용되기 위해 필요한 타당성과 공정성에 대한 과학적 근거는 아직 충분히 확보되지 못한 실정이다. 본 논문은 지금까지 진행되어온 머신러닝 기반 성격검사 연구(특히 소셜 미디어 텍스트 마이닝 방법을 중심으로)들을 분석하고 이를 토대로 향후 연구 방향을 제안한다. 검사의 타당성 및 공정성 확보를 위해서는 알고리즘 간 신뢰도, 검사-재검사 신뢰도, 집단간 타당성 차이 등에 대한 연구를 통해 과학적 기반을 확보해야 하며, 머신러닝 기술이 단순한 성격 예측의 도구가 아니라 인간의 성격에 대한 근본적 이해를 증진시키고 성격이론의 발전에 기여하도록 확장되어야 한다.

주요어 : 성격검사, 머신러닝, 소셜미디어 텍스트 마이닝, 타당성