Constructing a Psychometric Testbed for Fair Natural Language Processing

Ahmed Abbasi,[1,2]* **David Dobolyi,**[1,2] **John P. Lalor,**[1,2]
Richard Netemeyer,[3] **Kendall Smith,**[1,2] **Yi Yang**[4]
[1] Human-centered Analytics Lab, University of Notre Dame
[2] Department of IT, Analytics, and Operations, University of Notre Dame
[3] Center for Data Analytics, University of Virginia
[4] Department of Information Systems and Operations Management, HKUST
{aabbasi, ddobolyi, john.lalor, ksmith77}@nd.edu, rgn3p@virginia.edu, imyiyang@ust.hk

Abstract

Psychometric measures of ability, attitudes, perceptions, and beliefs are crucial for understanding user behavior in various contexts including health, security, e-commerce, and finance. Traditionally, psychometric dimensions have been measured and collected using survey-based methods. Inferring such constructs from user-generated text could allow timely, unobtrusive collection and analysis. In this work we construct a corpus for psychometric natural language processing (NLP) related to important dimensions such as trust, anxiety, numeracy, and literacy, in the health domain. We discuss our multi-step process to align user text with their survey-based response items and provide an overview of the resulting testbed, which encompasses survey-based psychometric measures and accompanying user-generated text from 8,502 respondents. Our testbed also encompasses self-reported demographic information, including race, sex, age, income, and education, allowing for measuring bias and benchmarking fairness of text classification methods. We report preliminary results on use of the text to predict/categorize users' survey response labels and on the fairness of these models. We also discuss the important implications of our work and resulting testbed for future NLP research on psychometrics and fairness.

1 Introduction

Psychometrics is the field of study concerned with the measurement of individuals' knowledge, abilities, attitudes, personality traits, and perceptions (Rust and Golombok, 2014). In social science research, psychometric dimensions are latent constructs that are known to be important antecedents, moderators, mediators, and consequents for important humanistic behaviors and outcomes. For example, constructs such as threat severity and response efficacy of protective mechanisms are critical psychometric measures of one's likelihood to avoid security threats (Zahedi et al., 2015). In behavioral health, psychometric dimensions such as health numeracy, subjective health literacy, trust in physicians, and anxiety visiting the doctor's office are known to effect various health and wellness outcomes such as future physician visits and all-around well-being (Netemeyer et al., 2020). In electronic commerce, satisfaction with a website's functional, information, and visual design are correlated with purchase propensity and customer loyalty (Cyr, 2008). Similarly, many individualized financial behaviors can be partially explained by financial literacy and psychological traits (Fernandes et al., 2014).

Given the importance of psychometric dimensions for understanding behaviors and outcomes in various domains, rigorous data collection protocols and best practices have been developed over the years (Netemeyer et al., 2003). The primary modes of collection involve surveys and interviews. While these techniques afford many benefits such as measurement control and robustness checks, they are not without their limitations. First, primary data collection facilitated through an administered survey can be time-consuming and invasive (often requiring 20-30 minutes of the respondents' time and attention). Second, such primary data collection cannot occur in real-time. Most surveys in field studies are conducted periodically at monthly or quarterly intervals. Third, while surveys are a rigorous form of data collection, they are limited in their ability to account for data/observations outside the predefined measurement framework. Effectively collecting and measuring relevant psychometric dimensions in a timely, unobtrusive, and open-ended manner could be invaluable in many real-world settings (Gefen and Larsen, 2017), including information retrieval and behavior modeling (Abbasi et al., 2015; Shing et al., 2020; Resnik et al., 2021).

*Authors listed alphabetically.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3748–3758
November 7–11, 2021.

In this paper we describe our efforts to construct a testbed for psychometric natural language processing (NLP). In the same vein as prior work on constructing language resources for sentiment, emotion, affect, and personality traits (Wiebe et al., 2005; Thelwall et al., 2010; Luyckx and Daelemans, 2008), and more recent work on modeling empathy and distress (Buechel et al., 2018; Abdul-Mageed et al., 2017), we describe our approach and resulting testbed related to psychometric dimensions such as trust, anxiety, literacy, and numeracy in the health context. Figure 1 presents a motivating example describing the goal of our work. Given a well-established survey-based scale for "trust in visiting the physician's office," how can we obtain a similar score based on user-generated text? Further, how do we ensure that our NLP-based scores are fair and unbiased?

The resulting testbed is comprised of user-generated text from 8,502 individuals for four key health-related psychometric dimensions of interest: trust in physicians, anxiety visiting the doctor's office, health numeracy, subjective health literacy. Our construction method and testbed contribute to the NLP language resource literature in the following ways:

- While psychometric dimensions such as sentiment, emotion, affect, and personality traits have garnered a fair amount of attention from the NLP community, there has been limited work on constructs like trust, anxiety, and perceptions of literacy.

- Given that psychometric analysis often entails user modeling that could involve analysis of text, survey-based responses (psychometric construct measures), and demographics, our testbed encompasses all three types of data.

- For each user, we capture text and gold-standard survey responses for four psychometric dimensions. The combination of four target dimensions, coupled with the aforementioned demographic and additional survey data affords opportunities for advanced text classification approaches such as multi-task learning and psychometric embeddings and encoders (Ahmad et al., 2020).

- By including text and demographics from diverse user populations, the testbed presents interesting opportunities for research on fairness in NLP models (Abbasi et al., 2018).

- While our efforts are geared towards psychometric dimensions in the health context, the method employed can be generalized to various contexts where psychometric dimensions are possible, practical, and valuable.[1]

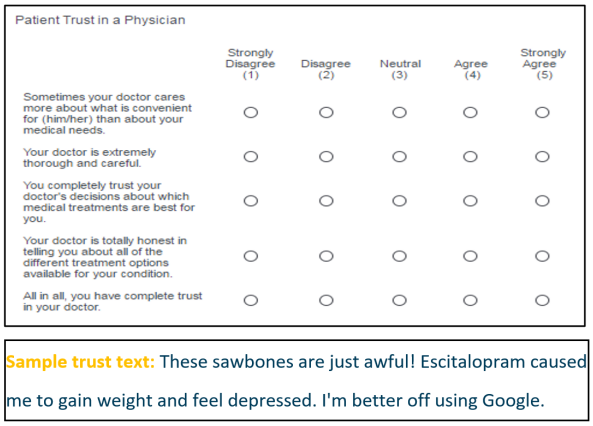


Figure 1: Illustration of survey and text data for a given psychometric dimension: "Trust in a Physician"

2 Testbed Construction Process

In this section we describe the process taken to construct our psychometric NLP testbed. The key steps included identifying relevant psychometric dimensions of interest, finding suitable survey-based items to operationalize our latent constructs, assessing different prompts for text equivalency questions, and testbed construction validation.

2.1 Identifying Key Psychometric Dimensions and Developing Survey Items

Given our focus on psychometrics in the healthcare context, we began by reviewing nearly 90 articles from the behavioral health literature (e.g., (Dugan et al., 2005; Schapira et al., 2014; Ciampa et al., 2010; Osborne et al., 2013; Netemeyer et al., 2020; Altin et al., 2014; Berkman et al., 2011). These articles all used survey-based methods to measure a set of core psychometric dimensions (i.e., latent constructs). Based on our literature review, we developed and tested a structural equation model that showed the relevant antecedent-consequent relations between various psychometric dimensions. Using this review and model, we further narrowed

[1]Code and data for this work are available at https://github.com/nd-hal/fair-psych-nlp

the consideration set down to four psychometric dimensions based on suitability of text-based response collection: trust in physicians (Dugan et al., 2005), anxiety visiting the doctor (Netemeyer et al., 2020), subjective literacy (Bishop et al., 2016), and objective health numeracy (Osborne et al., 2013). These four dimensions have also been found to be important antecedents or mediators for key health measures such as all-around perceptions of well-being and number of doctor visits. For instance, greater trust in physicians enhances well-being whereas one's perceptions of their health literacy increase such trust and also lower anxiety associated with visiting the doctor (Netemeyer et al., 2020). A critical step in survey-based psychometric research performed in the social sciences is development or inclusion of appropriate items to measure the latent constructs. Through our review of the literature, our own survey-based data collection, and statistical analysis (exploratory and confirmatory factor analysis), we identified a subset of items for each of these dimensions.

An overview of the four psychometrics dimensions and some of their related items is as follows. Note, the full items used appear in the readme file accompanying the dataset (included as part of the review process):

Health Literacy – In essence, health literacy (HL) is a subjective construct reflecting how much one thinks one knows about health and access to health-related information and providers (Osborne et al., 2013). Low HL has been associated with increased mortality, increased hospitalization, and poor adherence and self-maintenance to a host of chronic diseases such as diabetes, heart disease, and risk of stroke (Altin et al., 2014; Berkman et al., 2011; Osborne et al., 2013). Low HL has also been shown to be more prevalent among the elderly, lower income and education groups, and certain racial groups (Altin et al., 2014). In total, 10 HL items from three different scales were incorporated (Parker et al., 1995; Chinn and McCarthy, 2013; Bishop et al., 2016). Figure 2a shows examples of three of the items incorporated, which relate to one's perceptions of ability to understand hospital materials, process medical information, and comprehend medical conditions.

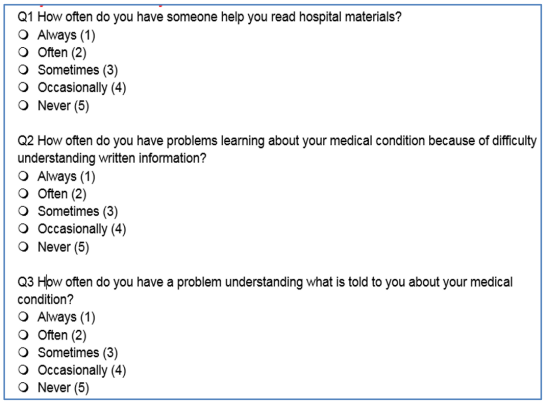
Health Numeracy – Conversely, health numeracy (HN) is an objective construct reflecting the ability to calculate, use, and understand numeric and quantitative concepts in the context of health issues (Schapira et al., 2014). HN has been associated with positive health outcomes such as the ability to understand dosage in medication and adherence to self-care diabetes treatment (Ciampa et al., 2010; Osborne et al., 2013). As with HL, lower HN scores are more prevalent among the elderly, lower income and education groups, and certain racial groups (Schapira et al., 2014). We incorporated two HN scales comprising 14 total items (Osborne et al., 2013; Schapira et al., 2014). Figure 2c depicts four item examples from one of the two scales utilized. As shown, these items are objective measures such as ability to count calories or read a thermometer.

Trust in Doctors – Perceptions of trust in physicians/doctors (TD) can have an important mediating role on health outcomes (Dugan et al., 2005). TD was measured using the well-validated 5 items proposed by (Dugan et al., 2005), depicted in Figure 1.

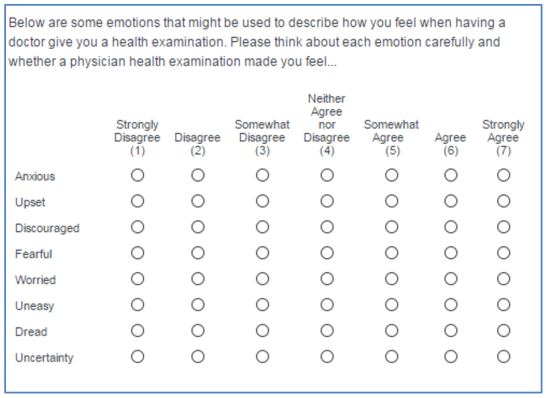
Anxiety Visiting Doctors – Anxiety when visiting the doctor's office (AV) is another strong potential mediator for health outcomes such as future doctor visits and wellness (Spielberger, 1989). Figure 2b shows the items used to measure AV. These focused on levels of anxiousness, worry, uncertainty, and uneasiness (Netemeyer et al., 2020).

2.2 Obtaining User-Generated Text

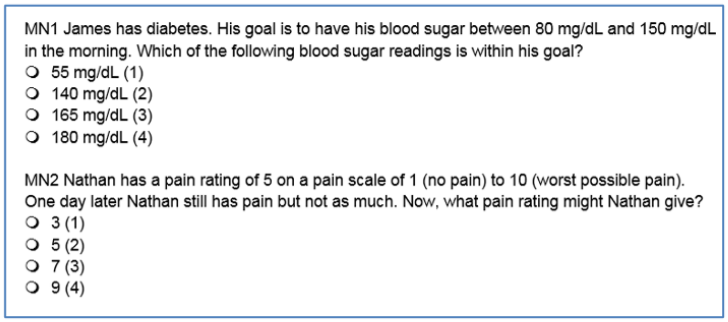
We used an iterative trial-and-error process to develop our "equivalent" user generated text related to the four aforementioned psychometric dimensions. The key design considerations were: (1) the placement of the text response box (e.g., same page as survey items or next page); (2) the questions/prompts used to elicit text responses. After several rounds of face validity checks and piloting with small sets of respondents, we ultimately arrived at a configuration where the survey items were used to prime respondents. We immediately followed these items with text questions that were tuned as part of our iterative process. The text-response questions yielded the best responses (i.e., in terms of alignment between text semantic orientation and survey items) when the questions were at the end of the survey item section for that particular psychometric dimension, appearing immediately at the bottom of the same/final page of survey items. Table 1 depicts the prompts or questions used to attain the user-generated text responses.

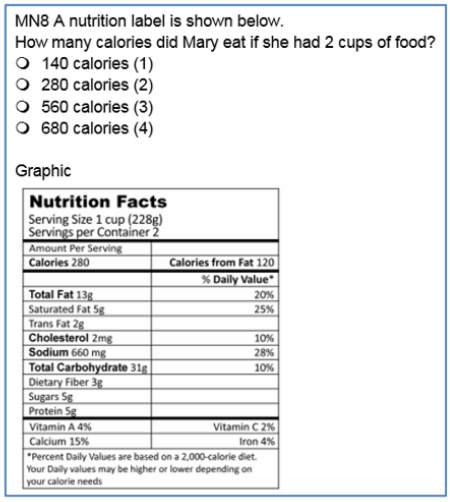


(a) Examples of health literacy survey items.



(b) Examples of anxiety in visiting the doctor survey items.





(c) Examples of health numeracy survey items.

Figure 2: Examples of survey items for three of the four psychometric constructs.

3 Testbed Results and Summary Statistics

Two rounds of data collection were performed using AMT and Qualtrics, respectively. In order to ensure high data quality, we followed best practices for crowd-sourced data collection including suitable compensation, validity checks, clear instructions, and manual inspection of the data (Buhrmester et al., 2011; Buechel et al., 2018). In each round, all responses were manually examined for quality assurance. A small proportion of responses were removed due to noisy text (e.g., failing to properly answer the questions), a failed validity check, or for responding too quickly (relative to the median response times). For both data collections, each participant was compensated five US dollars.

In the first round, we collected a total of 4,262 usable responses via Amazon Mechanical Turk (AMT). In order to attain a second, more diverse set of responses, Qualtrics was used to collect an additional 4,240 clean responses. Based on quantitative and qualitative assessment of the data, participants seemed engaged in the task and thoughtful in their responses - the mean and median response times were 32.7 and 24.1 minutes, respectively (which are in the same ballpark as (Buechel et al., 2018)).

Table 2 shows the consolidated testbed summary statistics. Each respondent provided a text response for each of the four psychometric dimensions (§2.1), in addition to survey responses to all dimension items as well as additional demographic and behavior questions. We received 33,882 total text responses from 8,502 users across the AMT and Qualtrics data collections (i.e., there were 126 missing responses, 0.37%). The mean text response lengths for the four psychometric dimensions were in the 179 to 226 character range. The AMT respondents tended to be more representative of the overall US population in terms of race, gender,

Psychometric Dimension	Question or Prompt
Anxiety visiting the doctor (AV)	In a few sentences, please describe what makes you most anxious or worried visiting the doctor's office.
Subjective health literacy (HL)	Regarding all the questions you just answered, to what degree do you feel you have capacity to obtain, process, and understand basic health information and services needed to make appropriate health decisions? Please explain you answer in a few sentences.
Trust in physicians (TD)	In a few sentences, please explain the reasons why you trust or distrust your primary care physician. If you do not have a primary care physician, please answer in regard to doctors in general.
Objective health numeracy (HN)	In a few sentences, please describe an experience in your life that demonstrated your knowledge of health or medical issues.

Table 1: Questions used to elicit user-generated text responses

and education. As noted earlier, one goal of the Qualtrics data collection was to garner a richer sample of responses from diverse populations in terms of race, sex, education, and income, to allow deeper exposition into issues of fairness of NLP models (Abbasi et al., 2018).

Characteristic	Description
Unique users	8,502
Text fields (per user)	Subjective literacy (SL) Objective numeracy (HN) Trust in physicians (TD) Anxiety visiting the doctor (AV)
Age	Mean: 40.5 Over 50: 28.7%
Race	65.5% white 28.7% black 5.8% other
Sex (female)	65.4%
Income (USD)	62.7% < $55K
Education	50.9% college grad or higher
Examples of other behavior/ psychometric dimensions	Usage of prescription drugs Presence of primary care physician Frequency of doctor visits Smoking and drinking frequency
Response times	32.7 minutes (mean) 24.1 minutes (median)
Mean response lengths (chars)	226 (SL), 221 (HN) 218 (TD), 179 (AV)

Table 2: Testbed summary statistics

The most critical survey response items in the data were the ones corresponding to the four psychometric dimensions. Following best practices from the social science literature, we constructed a single composite score for each of these dimensions by averaging across multi-item scales (Buechel et al., 2018). The scores were scaled to a 0-1 range. Figure 3 depicts the distribution of user responses for the four dimensions (HL, HN, TD, AV). We can see that for HL and TD, the responses followed a skewed Gaussian distribution. In contrast, AV, and to a lesser extent, HN, were more uniformly distributed.

Table 3 shows examples of psychometric scores

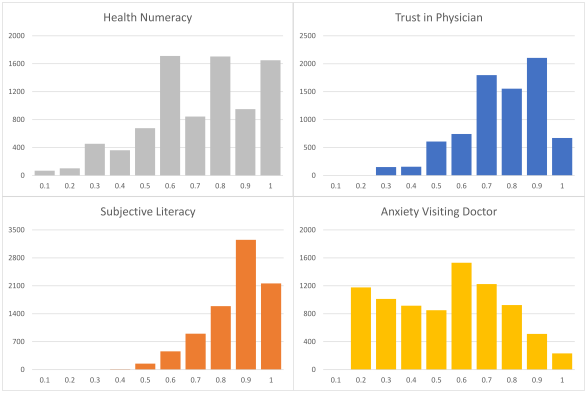


Figure 3: Distribution of response scores for psychometric dimensions

and accompanying text responses for the HL dimension. The scores were scaled from 0-1 based on the survey responses. The accompanying user text responses correspond to the two users' self-reported scores. The example illustrates the "alignment-oriented" objectives of testbed construction in this context (§2.2).

4 Modeling Literacy, Numeracy, Trust, and Anxiety

In order to evaluate the effectiveness of the constructed data set, we conducted regression and classification experiments to see how well various NLP models could predict survey-based "gold-standard" ratings using the free text responses. To ensure that each data point was evaluated, we used five-fold cross-validation. In each fold we used an 70/10/20 training/validation/testing split. Similar to prior studies (Buechel et al., 2018; Gibson et al., 2015), for the continuous prediction task, the dependent variable was the continuous 0-1 range labels for SL, HN, TD, and AV. For the classification task, we bifurcated our four dependent variables into high/low class labels (Gibson et al., 2015) by discretizing across the median values.

HL score	Text Response for Subjective Literacy Prompt
0.4667	I feel like with the terms and complicated medical lingo, I am not exactly sure what some of the meanings entail. Such as If I am diagnosed with a certain condition and need medication X, I don't know what that medication does, what the alternatives are, I don't even know how to pronounce some of these names. I feel like I am able to ask the doctors but can not fully grasp the magnitude of the situation without looking at the whole picture which is difficult to have someone explain to me in one visit.
0.9167	I think I have a fine capacity. I am able to coherently explain my concerns, and ask for aid if I need it. I am native in English, and know all my health issues and past surgeries and such. It isn't hard for me to do anything medical, and I am confident in making whatever medical decisions I need to make.

Table 3: Examples of survey-based scores and accompanying text responses for subjective health literacy

Model	Trust (TD)		Literacy (HL)		Numeracy (HN)		Anxiety (AV)	
	r	RMSE	*r*	RMSE	*r*	RMSE	*r*	RMSE
BERT	.619	.148	**.559**	**.094**	**.549**	**.192**	**.482**	**.204**
WordLSTM	**.621**	**.142**	.511	.101	.429	.228	.421	.211
WordCNN	.614	.146	.495	.127	.406	.219	.442	.210
FFNN	.604	.145	.489	.102	.356	.229	.394	.215
Regression	.535	.149	.331	.120	.206	.277	.332	.240

Table 4: Model performance for continuous prediction task (Pearson's *r* and RMSE)

Model	Trust (TD)		Literacy (HL)		Numeracy (HN)		Anxiety (AV)	
	AUC	F_1	AUC	F_1	AUC	F_1	AUC	F_1
BERT	**.845**	.739	**.798**	**.740**	**.776**	.706	**.723**	**.681**
WordLSTM	.810	**.772**	.754	.704	.704	.700	.689	.667
WordCNN	.806	.766	.760	.712	.717	**.708**	.684	.677
FFNN	.809	.770	.751	.695	.646	.707	.674	.656
Regression	.788	.712	.732	.667	.693	.640	.652	.601

Table 5: Model performance for binary classification task (AUC and F_1 score)

4.1 Model Regression and Classification Performance

We evaluated the data set against five NLP models:linear/logistic regression (LR), feed forward neural network (FFNN), word CNN, word LSTM, and BERT (Devlin et al., 2018). LR and FFNN were each run with a maximum of 50,000 word unigram, bigram, and trigram features. FFNN contained three dense layers each with 256 units, ReLU activation, L2 regularization of 0.001, each followed by a dropout layer with value of 0.5. Word CNNs and LSTMs both used the GloVe Common Crawl (840B token) 300 dimension word embeddings (Pennington et al., 2014). The word LSTM had two bidirectional layers with 128 units, each with dropout and recurrent dropout of 0.2, followed by a 64 unit dense layer. Following prior studies (Buechel et al., 2018; Majumder et al., 2017), the word CNN was a concatenation of three single convolutional layers of kernel size 1, 2, and 3 (i.e., to capture word unigram, bigram, and trigram level patterns), each with 256 filters and ReLU activation, followed by a global max pooling layer and a dense layer of 64 units. All three neural network models were trained using the Adam optimizer for 50 epochs with a learning rate of 0.0001 and a batch size of 32. For the regression task, the models used mean squared error for loss whereas for the classification task, they used binary cross entropy. BERT was run using the same architecture, optimization choices, and vocabulary as the BERT-base model (Devlin et al., 2018). Fine tuning was performed on our five-fold training data with mean squared error and cross entropy loss used for the regression and binary classification tasks, respectively.

For the regression tasks, consistent with prior research, BERT outperformed the LSTMs and CNNs, and the LSTMs attained better results than the feature-based FFNN and regression models (Table 4). Further, our highest Pearson's *r* values, in the 0.48 to 0.61 range, are on par with those attained for the well-established emotion intensity prediction problem (Mohammad and Bravo-Marquez, 2017; Strapparava and Mihalcea, 2007) and newer empathy and distress prediction tasks (Buechel et al., 2018; Gibson et al., 2015).

The binary classification task yielded similar results, with BERT outperforming the LSTM and CNN models in terms of AUC and F_1, and the LSTMs/CNNs in turn outperforming the FFNN and LR models (Table 5). Further, the best F_1 scores in the 0.68 to 0.77 range are comparable to results from prior studies classifying binary discretized labels (Gibson et al., 2015; Khanpour et al., 2017; Yates et al., 2017). The above regression and classification analysis underscores the effectiveness of our survey-text collection process and suggests that NLP-based modeling of psychometric dimensions such as literacy, numeracy, trust, and anxiety in health-related contexts might be possible and practical.

4.2 Model Fairness

As our data set includes rich demographic information, we can use it to evaluate the fairness of different NLP models (Friedler et al., 2019; Mehrabi et al., 2019; Blodgett et al., 2020). The data set includes five demographic variables: age, race, sex, income, and education (Table 2). While some prior NLP data sets have included user-level demographic information, it is rare, and to the best of our knowledge this is the first data set for NLP psychometrics with demographic information across these five variables. We believe the data set is well-aligned with recent calls for NLP bias research that examine the interplay between bias and harm in important application contexts (Blodgett et al., 2020).

To demonstrate an assessment of model fairness, we evaluated three of our NLP models (FFNN, WordCNN, and BERT) for fairness with regards to race. We binarized the race demographic variable such that "white" was the privileged class and "non-white" was the non-privileged class (Friedler et al., 2019). We calculated the disparate impact (DI) (Equation 1, Friedler et al. (2019); Mehrabi et al. (2019)) of positive predictions for all four of our dependent variables (left chart in Figure 4). DI is a useful metric here because appropriate positive prediction is necessary for possible interventions (e.g., referral to a health literacy specialist). $DI < 1$ indicates that there are fewer positive predictions for the non-privileged class than for the privileged class (e.g., fewer approved loan applications for non-whites relative to those that are white).

$$DI = \frac{p(\hat{Y} = 1|S \neq 1)}{p(\hat{Y} = 1|S = 1)} \quad (1)$$

For anxiety, subjective literacy, and trust in physicians, DI is generally close to 1, suggesting greater equity. For numeracy there is more variation across scores, in particular with respect to BERT. DI is much lower for BERT (less than 0.7) relative to FFNN (0.88) and WordCNN (1.0), suggesting that BERT's scoring of health numeracy text might be less fair. The BERT model is 30% less likely to assign a high numeracy score to non-white participants' text. We also evaluated the NLP models using the xAUC metric (Kallus and Zhou, 2019). xAUC considers the ranked nature of risk scores for potentially resource-constrained scenarios (e.g., physician availability). Specifically, we look at the difference between xAUC scores between groups:

$$\begin{aligned} \Delta\text{xAUC} &= p(R_1^a > R_0^b) - p(R_1^b > R_0^a) \\ &= p(R_1^b \leq R_0^a) - p(R_1^a \leq R_0^b) \end{aligned} \quad (2)$$

Positive ΔxAUC values indicate that group a's members in the positive class ($Y = 1$) have higher model scores than group b's members in the negative class ($Y = 0$). Looking at xAUC (right side of Figure 4), once again the values for numeracy when using the BERT and FFNN models indicate that there might be disparities between the privileged and non-privileged classes that are worth further investigation.

This analysis illustrates how the testbed can be used to model fairness. Further analysis could extend to the multi-class scenario for race, and may also be applied to the other demographic variables, making this a rich data set for future fair NLP research. In addition, because the gold-standard labels are continuous (e.g., a numeracy score), this data set can facilitate development of new fairness metrics that merge calibration (Pleiss et al., 2017) with class-label-focused fairness assessments such as DI and xAUC.

5 Related Work

Over the past thirty years, significant efforts have been made to develop a robust and burgeoning set of language resources for various linguistic and NLP tasks (Bowman et al., 2015; Guzmán et al., 2019). Gold-standard testbeds have been developed for sentiment analysis and emotion detection (Wiebe et al., 2005; Thelwall et al., 2010). Personality traits manifested in text have also received attention (Luyckx and Daelemans, 2008). More recent work has explored construction of corpora

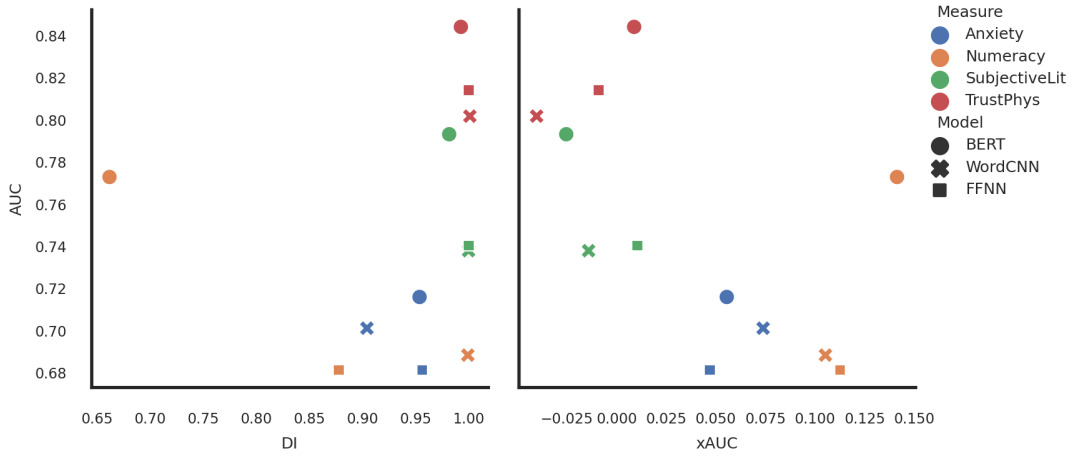


Figure 4: Plot of model performance (AUC) against fairness (DI on left, ΔxAUC on right)

for examining depression and cyberbullying, including annotating self-disclosures of personal information which may trigger bullying (Rakib and Soon, 2018), and testbeds for modeling empathy and distress (Buechel et al., 2018).

Given that psychometrics is concerned with measurement of attitudes, beliefs, perceptions, and personality traits, many of these aforementioned testbeds and avenues of language resource construction could be considered as focusing on psychometric dimensions (Ahmad et al., 2020). We build on this work by focusing on underexplored dimensions such as trust, anxiety, and perceptions of literacy in a health context. Moreover, rather than relying on independent annotation, we seek to utilize user-generated text that is captured along with self-reported survey-based responses for the psychometric dimensions of interest (Buechel et al., 2018). Hence, the text is accompanied by survey-based quantifications from the individuals that can serve as a gold-standard proxy of what we hope to measure by applying NLP methods.

This paper bridges the social science and NLP perspectives for testbed construction. Such work is aligned with recent efforts at the intersection of NLP and mental health such as psychological health prediction and suicide prevention (Lynn et al., 2018; Shing et al., 2020; Resnik et al., 2021). Consistent with prior work using self-reported survey-based items (Buechel et al., 2018) as gold-standard labels, we use supervised machine learning classification methods to demonstrate the viability of the approach – that is, to validate that the text samples captured can indeed serve as a reasonable proxy of the users' survey-based responses for the psychometric dimensions of interest. Further, our testbed also includes the users' survey-based responses to related psychometric dimensions, as well as demographic data. We use the latter to explore the fairness of our text classifiers - an important direction for current and future NLP research (Bender et al., 2021; Chang et al., 2019).

6 Conclusion

The results of our work have important implications for several stakeholder groups. NLP research focused on constructing novel empirical methods can use the constructed testbed to build new models for psychometric NLP. The inclusion of demographic, text, target psychometric, and secondary psychometric data in the testbed could allow development of rich deep learning architectures that incorporate user models (Ahmad et al., 2021), psychometric embeddings, structural equation model-based encoders, and multi-task learning across the four parallel target psychometric dimensions (Ahmad et al., 2020).

The unique multimodal nature of the data may also afford opportunities to better understand and study fairness in NLP models and methods (Blodgett et al., 2020). For each text utterance, the testbed encompasses gender, race, education levels, and income – all fields that are often the basis for bias in machine learning algorithms. While there is a rich and growing stream of research on bias and fairness in NLP, the examination of fairness in NLP using gold-standard demographic data (i.e., with known demographics of the authors) is to-date underexplored. This combination of downstream dependent variables and known demographics is an

important step towards analyzing NLP fairness issues in real-world social contexts with clear normative goals, while considering the lived experiences of the community members they affect (Blodgett et al., 2020; Taylor et al., 2018).

Finally, other teams developing language resources can adapt the process outlined to other domains such as security, e-commerce, finance, etc. We recognize that this is one of a handful of forays into rich psychometric NLP. Our hope is that future work can improve upon the methods and best practices for examining the interplay between survey-based constructs and their manifestations in user-generated text.

While we recognize that the questions asked and approach undertaken could be further enhanced, we believe this constitutes an important first step toward aligning survey items with user-generated text responses. As we show in the evaluation section, preliminary results from text classification tasks lend validity to the construction.

Any NLP-based approximation is likely to have measurement error due to the error of the text classifier trained to score the user text, as well as dissonance between a user's survey responses and text utterances. Nevertheless, the hope is that the ability to infer an imperfect yet reasonably accurate NLP-based measurement can still be advantageous as an alternative, complementary measure that can be derived unobtrusively in near real-time.

As noted, we believe the testbed and process have important implications for future NLP research that examines psychometrics and fairness as part of broader user modeling efforts.

Acknowledgements

The authors would like to thank the individuals who participated in our data collection process. This work was supported in part by U.S. NSF grants: IIS-2039915, IIS-1816504, BDS-1636933, and IIS-1553109.

References

Ahmed Abbasi, Raymond YK Lau, and Donald E Brown. 2015. Predicting behavior. *IEEE Intelligent Systems*, 30(3):35–43.

Ahmed Abbasi, Jingjing Li, Gari Clifford, and Herman Taylor. 2018. Make "fairness by design" part of machine learning. *Harvard Business Review*.

Muhammad Abdul-Mageed, Anneke Buffone, Hao Peng, Johannes Eichstaedt, and Lyle Ungar. 2017. Recognizing pathogenic empathy in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Faizan Ahmad, Ahmed Abbasi, Brent Kitchens, Donald A Adjeroh, and Daniel Zeng. 2021. Deep learning for adverse event detection from web search. *IEEE Transactions on Knowledge and Data Engineering*.

Faizan Ahmad, Ahmed Abbasi, Jingjing Li, David G Dobolyi, Richard G Netemeyer, Gari D Clifford, and Hsinchun Chen. 2020. A deep learning architecture for psychometric natural language processing. *ACM Transactions on Information Systems (TOIS)*, 38(1):1–29.

Sibel Vildan Altin, Isabelle Finke, Sibylle Kautz-Freimuth, and Stephanie Stock. 2014. The evolution of health literacy assessment tools: a systematic review. *BMC public health*, 14(1):1–13.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, and Karen Crotty. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*, 155(2):97–107.

Wendy Pechero Bishop, Simon J Craddock Lee, Celette Sugg Skinner, Tiffany M Jones, Katharine McCallister, and Jasmin A Tiro. 2016. Validity of single-item screening for limited health literacy in english and spanish speakers. *American journal of public health*, 106(5):889–892.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765.

M Buhrmester, T Kwang, and SD Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5.

Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference*

on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts.

Deborah Chinn and Catherine McCarthy. 2013. All aspects of health literacy scale (aahls): developing a tool to measure functional, communicative and critical health literacy in primary healthcare settings. *Patient education and counseling*, 90(2):247–253.

Philip J Ciampa, Chandra Y Osborn, Neeraja B Peterson, and Russell L Rothman. 2010. Patient numeracy, perceptions of provider communication, and colorectal cancer screening utilization. *Journal of health communication*, 15(sup3):157–168.

Dianne Cyr. 2008. Modeling web site design across cultures: relationships to trust, satisfaction, and e-loyalty. *Journal of management information systems*, 24(4):47–72.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elizabeth Dugan, Felicia Trachtenberg, and Mark A Hall. 2005. Development of abbreviated measures to assess patient trust in a physician, a health insurer, and the medical profession. *BMC health services research*, 5(1):1–7.

Daniel Fernandes, John G Lynch Jr, and Richard G Netemeyer. 2014. Financial literacy, financial education, and downstream financial behaviors. *Management Science*, 60(8):1861–1883.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

David Gefen and Kai R Larsen. 2017. Controlling for lexical closeness in survey research: A demonstration on the technology acceptance model. *Journal of the Association for Information Systems*, 18(10):1.

James Gibson, Nikolaos Malandrakis, Francisco Romero, David C Atkins, and Shrikanth S Narayanan. 2015. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *Sixteenth annual conference of the international speech communication association*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Nathan Kallus and Angela Zhou. 2019. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In *Advances in neural information processing systems*.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251.

Kim Luyckx and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *LREC*.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, LA. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Richard G Netemeyer, William O Bearden, and Subhash Sharma. 2003. *Scaling procedures: Issues and applications*. Sage Publications.

Richard G Netemeyer, David G Dobolyi, Ahmed Abbasi, Gari Clifford, and Herman Taylor. 2020. Health literacy, health numeracy, and trust in doctor: Effects on key patient health outcomes. *Journal of Consumer Affairs*, 54(1):3–42.

Richard H Osborne, Roy W Batterham, Gerald R Elsworth, Melanie Hawkins, and Rachelle Buchbinder. 2013. The grounded psychometric development and initial validation of the health literacy questionnaire (hlq). *BMC public health*, 13(1):1–17.

Ruth M Parker, David W Baker, Mark V Williams, and Joanne R Nurss. 1995. The test of functional health literacy in adults. *Journal of general internal medicine*, 10(10):537–541.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5684–5693.

Tazeek Bin Abdur Rakib and Lay-Ki Soon. 2018. Using the reddit corpus for cyberbully detection. In *Asian conference on intelligent information and database systems*, pages 180–189. Springer.

Philip Resnik, April Foreman, Michelle Kuchuk, Katherine Musacchio Schafer, and Beau Pinkham. 2021. Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior*, 51(1):88–96.

John Rust and Susan Golombok. 2014. *Modern psychometrics: The science of psychological assessment*. Routledge.

Marilyn M Schapira, Cindy M Walker, Tamara Miller, Kathlyn E Fletcher, Pamela S Ganschow, Elizabeth A Jacobs, Diana Imbert, Maria O'Connell, and Joan M Neuner. 2014. Development and validation of the numeracy understanding in medicine instrument short form. *Journal of health communication*, 19(sup2):240–253.

Han-Chin Shing, Philip Resnik, and Douglas Oard. 2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137, Online. Association for Computational Linguistics.

Charles Donald Spielberger. 1989. *State-trait anxiety inventory: a comprehensive bibliography*. Consulting Psychologists Press.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Herman A Taylor, Frances Henderson, Ahmed Abbasi, and Gari Clifford. 2018. Cardiovascular disease in african americans: innovative community engagement for research recruitment and impact. *American Journal of Kidney Diseases*, 72(5):S43–S46.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978.

Fatemeh Mariam Zahedi, Ahmed Abbasi, and Yan Chen. 2015. Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance. *Journal of the Association for Information Systems*, 16(6):2.