PyNDA: Deep Learning for Psychometric Natural Language Processing

Jingjing Li
jl9rf@comm.virginia.edu
University of Virginia

Ahmed Abbasi
abbasi@comm.virginia.edu
University of Virginia

Faizan Ahmad
faizann288@gmail.com
University of Virginia

Hsinchun Chen
hchen@eller.arizona.edu
University of Arizona

1 Introduction

PSYCHOMETRICS is concerned with the measurement of knowledge, ability, attitudes, and personality traits. With the increased importance of predictive analytics at the micro-level [1], including prediction of individuals' behaviors [2], accurate and timely measurement of psychometrics has become of paramount importance. In health settings, psychometric measures, including health numeracy, subjective literacy, and perceptions of trust and anxiety related to physicians, have been shown to have a profound impact on various health and wellness outcomes such as future doctors' visits and all-around well-being [6], [7], [8]. Hence, accurately and efficiently measuring psychometrics can have positive implications for many behavior prediction tasks [42], [43], [48], [49], [51], [52].

Psychometric data collection efforts have traditionally relied on survey-based methods administered on a monthly or quarterly basis. Effectively collecting and measuring relevant constructs in a timely and unobtrusive manner has proven elusive in real-world settings [9]. In recent years, machine-learning methods for natural language processing (NLP) have been successfully applied to certain psychometric dimensions such as sentiment and emotion [10]. Such NLP techniques, which analyze user-generated text and automatically score them along the target variable, afford opportunities for real-time, passive monitoring and measurement. However, several gaps and challenges remain:

- *Many rich psychometric dimensions remain underexplored*: Whereas numerous NLP methods have been proposed for sentiment and emotion, other aspects such as attitudes, perceptions, and characteristics have received limited attention. It is unclear how effectively NLP methods can tackle these novel dimensions [50].
- *User-centric versus task-centric modeling*: Most prior NLP classification objectives and data sets have been arranged around a given task (e.g., sentiment polarity). Psychometric dimensions such as attitudes and perceptions are very individualized, with multiple inter-related target variables of interest associated with each person. There is an opportunity for psychometric NLP methods to incorporate provisions for user-centric modeling [44], [47].
- *Demographic-sensitive modeling*: Factors such as age, race, gender, and education can have a profound impact on various psychometric measures (e.g., literacy, trust, anxiety) [5]. These differences can be amplified in user-generated text [11]. Several recent studies suggest that machine learning models that fail to properly control for demographics are prone to inaccurate generalizations [12], [46]. Psychometric NLP methods that are accurate across diverse demographic populations are a necessary and understudied research area [45].
- *Paucity of available text*: Several recent NLP studies have examined "short-text" contexts such as Twitter [13] and news articles [14]. User-generated text associated with psychometrics often appears in similarly sparse environments such as comment boxes, text messages, and microblogs, necessitating methods capable of learning patterns from limited linguistic cues.

In order to address these gaps, we propose a novel deep learning architecture for psychometric NLP. Our architecture incorporates provisions to address the aforementioned issues, including novel representation and demographic embeddings and a structural equation modeling (SEM) encoder, coupled with a robust multi-task learning method. The proposed architecture was evaluated on a rich health test bed encompassing three data sets comprised of pertinent psychometric dimensions — such as health numeracy, literacy, trust, anxiety, and drug experiences — related to a set of demographically diverse users. The results reveal that the proposed architecture is able to garner markedly better classification accuracy, precision, and recall rates across psychometric dimensions, relative to baseline and benchmark machine learning NLP methods. Ablation analysis shows that each component significantly contributes to overall performance, thereby underscoring the efficacy of the proposed architecture.

2 RELATED WORK

Based on our review of relevant literature on feature-based NLP classifier [31] [17], Recurrent Neural Network (RNN) [31], Convolutional neural networks (CNN) [32], hybrid deep learning architectures [33], and multi-task adversarial learning [35], we have identified three major research gaps. First, although psychometric dimensions such as sentiment and emotion have been studied extensively, there has been limited focus on other rich psychometric dimensions such as trust, anxiety and literacy. Research examining such psychometric dimensions is of theoretical and practical importance. For instance, effectively capturing such psychometric dimensions necessitates consideration of user-centric modeling techniques capable of considering inter-related dimensions in unison, as well as demographic-sensitive modeling. Second, little work has been done to fuse the rich linguistic resources, methods, and domain knowledge developed in the feature-based NLP classification literature with novel deep learning architectures. Given the complexity of psychometric utterances and paucity of available text, such fusion could facilitate enhanced accuracy by leveraging rich linguistic feature representations in concert with robust deep learning schemes. Third, hybrid deep learning architectures encompassing CNNs, LSTMs, and multi-task learning mechanisms have been underexplored. Prior work suggests these approaches offer complementary benefits such as pattern detection from local features, consideration of long-term dependencies, and inclusion of the interplay between closely related user-level psychometric dimensions. In the ensuing section, we propose an architecture expressly designed to address these gaps.

3 PROPOSED DEEP LEARNING ARCHITECTURE

Figure 1 depicts our proposed PyNDA **P**s**y**chometric **N**LP **D**eep Learning **A**rchitecture, which encompasses four base neural nets that are fused via a concatenation layer that feeds into dense layers and also leverages a novel multi-task learning mechanism. Each component of the architecture is intended to address the aforementioned research gaps, thereby resulting in enhanced text classification capabilities for psychometric dimensions:

- A *character embedding* convolutional neural network (CNN) for capturing fundamental spatial syntactic patterns in user-generated texts at the character and prefix, suffix, and root levels.
- A bi-directional long short-term memory (Bi-LSTM) recursive neural network that uses a novel underlying parallel *representation embedding* that encompasses an array of topic, sentiment, emotion, and syntactic linguistic representations. This embedding leverages feature subsumption methods capable of ingesting large, diverse feature spaces and refining them into a small set of rich attributes.
- A second Bi-LSTM that incorporates a novel *demographic embedding* scheme intended to better capture nuances and norms inherent across different gender, race, and age segments.
- A structural equation model *(SEM) Encoder* that allows inclusion of related "secondary" attitude and behavior information to allow superior classification of key target psychometric dimensions.
- A novel multi-task learning mechanism that enables better inclusion of joint information between related target psychometric dimensions.

In the remainder of the section, we describe each component of the proposed architecture.

3.1 Character Embedding

In order to consider the morphological patterns (e.g. prefix, suffix and misspelling) of the input text, we build a character-level embedding using a convolutional neural network. The input for the character embedding is a sequence of encoded characters. Each character is represented as a one-hot (or one-over-l) vector $g(x) \in [1, l] \to \mathbb{R}$, where l is the size of the alphabet. The alphabet used in our model consists of 70 characters, including 26 English letters, 10 digits, 33 other characters, and the new line character. The convolutional kernel function is defined as $f(x) \in [1, k] \to \mathbb{R}$, where k is the size of the filters. Given the stride of d we can get the convolution $h(y) \in \left[1, \left\lfloor\frac{l-k+1}{d}\right\rfloor\right] \to \mathbb{R}$ between $f(x)$ and $g(x)$ as follows:

$$h(y) = \sum_{x=1}^{k} f(x) \bullet \mathrm{g}(\mathrm{y} \bullet \mathrm{d} - \mathrm{x} + \mathrm{c}) \qquad (1)$$

where c = k − d + 1 is an offset constant. This convolutional layer is later connected to a max-pooling layer, defined as:

$$h(y)_{\max\_pooling} = max_{x=1}^{k} g(\mathrm{y} \bullet \mathrm{d} - \mathrm{x} + \mathrm{c}) \qquad (2)$$

The embedding process uses two convolutional layers, each followed by a max pooling layer. The resulting embedding is fed into two fully-connected layers, which are then concatenated with layers from other embeddings for the finally psychometric variable classification.

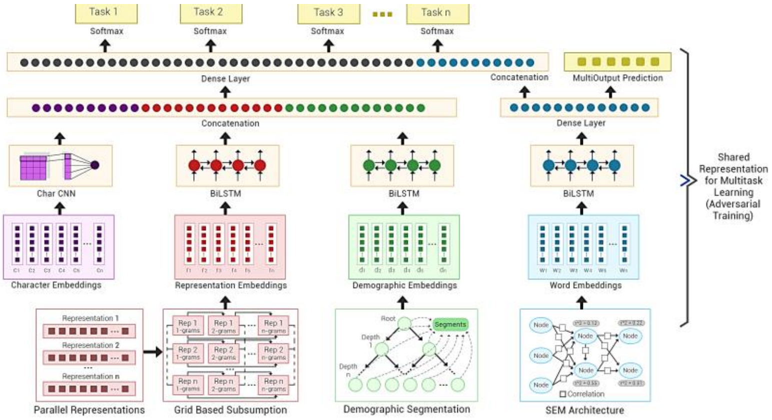

Fig. 1. PyNDA - proposed psychometric NLP deep learning architecture.

3.2 Representation Embedding

Examination of rich psychometric dimensions pertaining to diverse user demographics could pose challenges for deep learning methods, particularly in situations involving limited user-generated text. Recent work has shown that rich feature-based methods can often attain text classification performance levels that are comparable to simple deep learning architectures [15], whereas combining the two can often yield enhanced performance [16]. Accordingly, we propose a novel representation embedding that utilizes a rich array of parallel feature representations that capture a bevy of semantic and syntactic information at varying granularities, coupled with grid-based subsumption. The main intuition behind the proposed embedding is similar to standard word embeddings: create a lower dimensional feature space that captures key patterns. However, as we later demonstrate empirically, the representation embedding is particularly well-suited for psychometric NLP, providing strong discriminatory potential.

3.2.1 Parallel Representations

The major parallel representations incorporated for the input text include semantic and syntactic category. The semantic category encompassed topic, sentiment, and emotion related representations. The syntactic representations incorporated were parts of speech (POS) tags, words combined with POS tags, misspellings, and legomena. The combination of words with their respective parts of speech were included as an additional layer of word disambiguation [17].

3.2.2 Grid-Based Subsumption (GBS)

Although parallel representations can allow inclusion of rich linguistic representations at varying granularities, it also creates potential for inclusion of noise, redundancy, and irrelevant information. Accordingly, prior studies have proposed the use of subsumption methods to rectify this concern: feature space reduction techniques specifically crafted for natural language data [17] and [18]. However, prior methods use small, pre-defined subsumption mechanisms that are not scalable or extensible to large, dynamic feature spaces (e.g., [17], [18]). In order to overcome these limitations, we propose a novel grid-based subsumption (GBS) method well-suited for "winnowing the wheat from the chaff" atop our rich parallel representations. GBS uses a four-stage algorithm.

Stage 1 of GBS is mostly consistent with prior subsumption methods [17], [18], where only higher-order n-grams with enhanced discriminatory potential are retained over their lower-order n-gram feature counterparts within the same-representation. Given the set of m representations R = $\{r_1, r_2,\ldots r_m\}$, where each r_x signifies a parallel representation (e.g., word), we extract all n-gram features such that any f_{ijx} element in feature set F represents the ith feature in n-gram category j for representation r_x, and f_{ijx} is initially weighted as follows:

$$w(f_{ijx}) = \max_{c_a,c_b}\left(p(f_{ijx}|c_a)\log\left(\frac{p(f_{ijx}|c_a)}{p(f_{ijx}|c_b)}\right)\right) + s(f_{ijx}) \quad (3)$$

where c_a and c_b are amongst the set of C class labels, $c_a \neq c_b$, y is one of the d tokens in f_{ijx} with w possible word senses, and function s is the semantic orientation score, computed as the difference between the positive and negative polarity scores for sense q of token f_{ijxy} in SentiWordNet:

$$s(f_{ijx}) = \sum_{y=1}^{d}\sum_{q=1}^{w}\left(\frac{\mathrm{pos}(f_{ijxy}, q) - \mathrm{neg}(f_{ijxy}, q)}{dw}\right) \quad (4)$$

The first part of the weighting equation considers the discriminatory potential of the feature based on its log-likelihood ratio, whereas the second part factors in the semantic orientation to ensure that features with opposing orientation (e.g., "like" versus "don't like") are differentiated in terms of overall weights and when making subsumption decisions. Once features are weighted, the within representation r_x subsumption is performed as follows. Each n-gram feature f_{ijx} with $w(f_{ijx}) > 0$ is compared against each lower-order n-gram feature f_{uvx}, where $v < j$, $w(f_{uvx}) > 0$, and f_{uvx} contains some subsequence of tokens from f_{ijx}. If $c(f_{ijx}) = c(f_{uvx})$, where:

$$c(f_{ijx}) = \arg\max_{c_a,c_b}\left(p(f_{ijx}|c_a)\log\left(\frac{p(f_{ijx}|c_a)}{p(f_{ijx}|c_b)}\right)\right) \quad (5)$$

Then we determine whether to subsume the higher order n-gram as follows, where t is a subsumption threshold:

$$w(f_{ijx}) = \begin{cases} 0, & \text{if } w(f_{ijx}) \leq w(f_{uvx}) + t \\ w(f_{ijx}), & \text{otherwise} \end{cases} \quad (6)$$

Stage 2 entails cross-representation subsumption. Prior studies have relied on manually crafted subsumption graphs encompassing predefined representations and relation links (e.g., [17], [18]). In order to make the subsumption process more dynamic and extensible across an array of novel psychometric dimensions, we propose a graph construction approach. For each unique pair of representations r_x and r_z in R, let A and B signify randomly selected subsets of m features from these representations where each $f_{ijx} \in A$ and $f_{uvz} \in B$ is such that $j,v = 1$ (i.e., only unigram features). Since representations vary with respect to feature frequency and co-occurrence patterns, it is important to factor in such nuances by considering within category similarities when making cross-category comparisons. We use k-Means clustering to find the ideal partition over the $2m$ feature sample encompassing all elements in $A \cup B$. With $k = 2$, this results in $G = \{g_1, g_2\}$ clusters. A link is formed between r_x and r_z if the cross-cluster entropy reduction ratio attributable to representation affiliation information is below a certain threshold:

$$L(r_x, r_z) = \begin{cases} 1, & \text{if } \dfrac{H(G|r)}{H(G)} \leq l \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where $H(G)$ is the entropy across clusters and $H(G|r)$ is:

$$H(G|r) = -\sum_{r\in\{r_x,r_z\}} P(r)\sum_{g\in G} P(g|r)\log_2 P(g|r) \quad (8)$$

In Stage 3, once links are formed between representations as described in equation 7, cross-representation subsumption between any pair of r_x and r_z where $L(r_x, r_z) = 1$ is performed similarly to the approach described in equations 5 and 6 of Stage 1. Since the links are bidirectional, $L(r_x, r_z) = L(r_z, r_x)$. Hence, two-way comparisons are made, where each remaining f_{ijx} with $w(f_{ijx}) > 0$ in r_x is compared against each lower-order n-gram feature f_{uvz} where $v < j$, $w(f_{uvz}) > 0$, and f_{uvz} contains some subsequence of tokens from f_{ijx}, and then, each remaining f_{uvz} with $w(f_{uvz}) > 0$ in r_z is compared against its lower-order n-gram counterparts in r_x meeting the same criteria.

Finally, in Stage 4, we account for highly correlated non-subsuming cross-representation features. For each pair of r_x and r_z where $L(r_x, r_z) = 1$, each remaining f_{ijx} with $w(f_{ijx}) > 0$ in r_x is compared against all remaining f_{uvz} in r_z with weight greater than 0, where $j = v$. If the correlation between f_{ijx} and f_{uvz} is greater than threshold p, $w(f_{ijx}) = 0$.

3.2.3 Embedding and BiLSTM

For each representation, we use word2vec to learn an l-sized embedding vector for each token in that representation's data. However, only tokens with $w(f_{ijx}) > 0$ are included. For all other tokens, the embedding vector is replaced with a vector comprised of 0s. This embedding is then fed into a Bi-LSTM layer to learn the sequential dependency among words. The Bi-LSTM is later concatenated with hidden features of other embeddings as well as a softmax trained on weighted vectors where binary presence of "1" is replaced with $w(f_{ijx})$ for each token in the text.

3.3 Demographic Embedding

Demographics can have a profound impact on individuals' language usage tendencies and psychometric characteristics [4]. We build a novel demographic word-embedding to capture nuances and norms inherent to different demographic segments. More specifically, the demographic embedding identifies segments with the greatest entropy for a target psychometric dimension such that modeling within versus across such demographics may alleviate systematic bias [9] and enhance classification potential.

The first task is to identify demographic variables that significantly affect the psychometric dimensions of interest. We use a decision tree model to accomplish this task. Given a data set $\{a_1, a_2, \dots, a_M, C\}$, where $A = \{a_1, \dots, a_m, \dots a_M\}$ is the set of input demographic attributes and $C = \{c_1, \dots, c_N\}$ is the target psychometric classes, the decision tree partitions this dataset S into subsets using "nodes" according to input attribute a_m at certain splitting values $v \in V(a_m)$. $V(a_m)$ is the set of all possible values for attribute a_m. The goal is to create tree subdivisions that provide discriminatory potential for a given target class c_n. In this study, we use the entropy-based information gain metric as the node selection metric.

Given a target class C with possible values $\{c_1, c_2, \dots, c_m\}$ and probability mass function $P(C)$, the entropy H for the target class is defined as:

$$H\,(C) = -\sum_{i=1}^{m} P(c_i)\log_2 P(c_i) \qquad (17)$$

The information gain measures the reduction of entropy for target classes when further splitting the dataset by a new input attributea_m. Discretization is applied to continuous attributes before calculating the information gain. Specifically, the information gain of introducing an attribute $a_m$is defined as:

$$G(C, a_m) = H(C) - H(C|a_m = v_m) \qquad (18)$$

where *H(C)* is the entropy of the class label C and the second term is the expected entropy after the dataset is partitioned using attribute a_m at value v_m.

For the demographic embedding, we build two types of decision trees. The first type utilizes all demographic variables, termed as "global tree" T_g. The second type consists of a collection of "local trees" T_{lj}, each of which excludes one among the demographic variables. In the same spirit as the random forest algorithm [28], these local trees build on a random subset of input attributes to alleviate the possible dependency on a few dominant attributes. In order to be computationally feasible, we employ a binary tree structure and use depth parameter $d = 1,2, \dots$ D to control the tree size. The demographic trees are formulated as follows:

$$\mathrm{T_g} = \{a_m = v_m |\ a \in A, ht(T_g) = d\} \qquad (19)$$

$$\mathrm{T_{lj}} = \{a_m = v_m \mid a \in A \setminus \{a_i\}, ht(T_{lj}) = d\} \qquad (20)$$

where $ht()$ is the height function of the tree. The most prominent demographic conditions affecting the psychometric classes are selected based on node score *I*:

$$I(a_m = v_m) = \frac{NA(a_m = v_m)}{H(C|a_m = v_m)} + \frac{N(a_m = v_m)}{N(S)} \qquad (21)$$

where $a_m = v_m$ is the node representing a condition defined by an attribute a_m and its splitting value v_m (e.g., Age = 35); $NA(a_m = v_m)$ is the average of the accuracies of all the leaves underneath this node; $H(C|a_m = v_m)$ is the entropy with regards to class label for this node; $N(a_m = v_m)$ is the number of data points belong to this node; and $N(S)$ is the total number of data points in the data set.

The final set of demographic conditions M incorporated include the root node of the global tree and the top *K-1* nodes (ranked by node score *I*) from the local trees:

$$M = \{a_0 = v_0 \mid T_g \} \cup \{a_m = v_m \mid T_{li}, I(a_m = v_m) \in r_{tl}(I_1, I_2, \dots, I_{K-1})\} \tag{22}$$

where $a_0 = v_0 \mid T_g$ is the root node condition for the global tree and $r_{tl}(I_1, I_2, \dots, I_{K-1})$ is the top *K*-1 node scores for the local trees. The demographic embedding leverages this information as follows:

(1) Let m_k represent one of the *K* elements in *M*. For each m_k, we identify a subset of individuals satisfying that condition in the training set and construct a sub-corpus comprising text only belonging to those individuals. We use word2vec to learn an l-sized word embedding vector for each word j in the sub-corpus m_k such that $w_{kj} = (w_{kj1}, w_{kj2}, \dots w_{kjl})$. We also train a general word embedding across each word in the entire training set $w_j = (w_{j1}, w_{j2}, \dots w_{jl})$.

(2) For each individual u^i, we can identify the subset $M_s = \{m_1, m_2, \dots, m_s | m \in M\}$ of demographic conditions applicable to that user. Following the average embedding idea [36], the demographic embedding weight wd_{ij} for word j appearing in a text instance associated with individual u^i is defined as the weighted average of node score I_w and the node-specific word embedding w_{sj}:

$$wd_{ij} = \begin{cases} \dfrac{\left(\sum_s^{|M_s|} I_w * w_{sj}\right)}{|M_s|}, & if\ |M_s| > 0 \\ wd_{ij} & , if\ |M_s| = 0 \end{cases} \tag{23}$$

3.4 Structural Equation Model (SEM) Encoder

Psychometric dimensions are inherently correlated. For example, a patient with high anxiety associated with seeing a physician may also have low self-esteem [21]. In order to incorporate such secondary psychometric dimension information in PyNDA, we propose a novel Structural Equation Model (SEM) encoder. The underlying intuition behind our encoder is similar to the feature augmentation idea commonly used in multi-task learning, which and has been shown to offer significant performance lifts. Similarly, as illustrated in the ablation analysis in Section 4, our SEM encoder significantly enhances performance for classification of psychometric dimensions. Details are as follows.

SEM is a general multivariate statistical modeling technique to depict and test relationships among variables related to psychometric measures [29]. It models the psychometric dimensions as latent variables and discovers their most suitable relationships based on data. The SEM encoder aims to incorporate these multivariate, structured correlations between psychometric dimensions into PyNDA. Specifically, we build a series of SEM models for a given target psychometric dimension of interest along with other dimensions potentially affecting it. Let *S* represents a set of SEM models for a target psychometric dimension. Each model *G_i* in *S* can be considered a directed graph containing latent variables (or nodes) and directed links, arranged in a linear sequence with *K* levels and *J* nodes for each level such that node n_{kj} at level $k \geq 1$ connects to each n_{k+1j} in the next level. P is the path coefficient from an antecedent n_{kj} leading to a consequent variable n_{k+1j} with variance R^2 across all of its inbound antecedent links from level *k*. The target psychometric dimension only appears in level $k \geq 2$ to ensure it has antecedent variables and valid P and R^2. For each G_i, we can obtain the model fit indices CFI, TLI and RMSEA. In order to include a balanced model fit measure, we use $MF = (CFI + TLI + (1 - RMSEA))/3$ to depict the average model fit indices. For each non-target variable $v \in V$, we find a subset S' of all the SEM models containing them. We use a scoring function which weights path coefficients and model fit indices equally to summarize the relevance of any *v* to the target variable:

$$w(v) = \frac{1}{|S'|}\left|\sum_{S'} P\right| + \frac{1}{2|S'|}\left(\sum_{S'} R^2 + \sum_{S'} MF\right) \tag{24}$$

Finally, for each target variable we can derive the top *K* from *V* based on *w(v)* values. In order to avoid future leaks, we assume that *V* is unknown for test instances and must be predicted. Hence, a model is built on the training data to jointly score each selected *v*. This is done using a standard word embedding, followed by a Bi-LSTM layer and a fully-connected dense layer to classify the selected independent variables. The learned dense layer is then directly concatenated with the ones yielded by other embeddings to classify the target psychometric dimensions of interest.

3.5 Structural Multi-task Learning

Given the user-centric nature of psychometric analysis, structural relationships among target psychometric measures provide a unique opportunity for multitask learning (MTL). For example, if "trust in doctors" and "anxiety of seeing physicians" are correlated, we can share their input text features and jointly train the two classifiers together to augment the feature set for the current task.

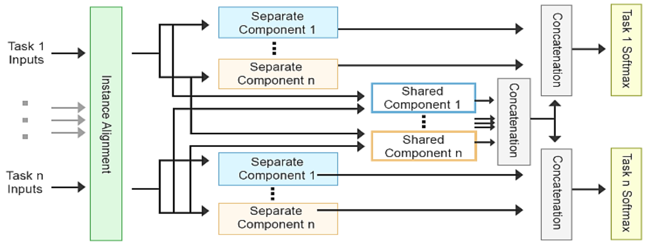

Fig. 2. Structural multi-task learning for psychometric dimensions

Figure 2 presents our proposed MTL approach. Suppose we have four target variables of interest, and wish to share features amongst them. Following [40], we create "separate LSTMs" comprising task-specific features and a single "shared LSTM" as a cross-task representation that reflects common patterns and cues across the different classification tasks. We jointly train these classifiers to allow feature sharing. In order to maintain orthogonality between shared and separate representations [40], adversarial training is used to optimize the purity of the shared representations. The idea is to build two agents, generator and discriminator, to combat one another. The generator tries to generate the purest shared feature set, while the discriminator attempts to distinguish the shared features into specific tasks. This tension results in convergence when the discriminator is no longer able to perform such differentiation, thereby resulting in better feature sharing. Collectively, this is accomplished via the loss function L = $L_{Task} + \lambda L_{Adv} + \gamma L_{Diff}$. As depicted in Figure 4, we extend this idea for our context in two ways. As later demonstrated in the ablation analysis, structural multitask learning with adversarial training provides an additional performance lift for our psychometric extraction tasks.

4 Testbed

In order to evaluate the proposed PyNDA architecture, an extensive research testbed was constructed, comprising three data sets and eleven total classification tasks. While psychometrics are known to be important in various application domains including security and e-commerce, in this study we focused on the health domain. The first two datasets encompassed four important psychometric dimensions known to be predictive of health outcomes.

1) Health Literacy (HL) – In essence, HL is a subjective construct reflecting how much one thinks one knows about health [19]. 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 [19].

2) Health Numeracy (HN) – 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. HN has been associated with outcomes such as the ability to understand dosage in medication and adherence to treatment [20].

3) Trust in Doctors (TD) – Perceptions of trust in physicians/doctors (TD) can have an important mediating role on health outcomes [21].

4) Anxiety Visiting Doctors (AV) – Anxiety when visiting the doctor's office is another strong mediator for health outcomes such as future doctor's visits and wellness [22].

For each of the four aforementioned psychometric dimensions (HL, HN, TD, and AV), well-established survey items have been developed in the literature. These items can be used to compute individuals' scores on a fixed continuous scale (e.g., 1-10). In order to construct our user-generated text datasets, we developed equivalent free response questions with accompanying text boxes that immediately followed the survey items. These questions were validated through pre-testing and were found to nicely represent the target variable for each users' collected text.

Table 1 summarizes the three datasets and relatedg classification tasks incorporated in our testbed. Consistent with several prior psychometric and NLP studies, for our first dataset we used Amazon Mechanical Turk (AMT) since it is considered somewhat representative of the broader Internet population. Each respondent provided quantitative and text responses for all four target dimensions of interest, some additional secondary dimensions, plus demographics such as age, gender, race, income, etc [3].

TABLE 1
SUMMARY OF TESTBED: THREE DATASETS AND 11 TASKS

Characteristics	**AMT**	**Qualtrics**	**AskAPatient**
Text Instances	4,262	4,240	138,998
Classification Tasks	Subjective literacy (HL) Health numeracy (HN) Trust in doctors (TD) Anxiety in visiting (AV)		Drug expert. Age Gender
Demographics:			
Race	81.2% white 7.4% black	50% white 50% black	unavailable
Age (mean)	37.4	45.6	39.9
Gender (male)	48.3%	24.2%	29.4%
Income (USD)	62% < $55K	67% < $55K	unavailable
Education (college grads)	44.6%	32.1%	unavailable

Since many health outcomes disproportionately impact health disparate populations, for our second dataset we used Qualtrics to collect 4,240 usable responses. Using the same survey instrument employed for AMT, the Qualtrics dataset was split evenly between Caucasian and African American respondents. In addition to race, individuals in this dataset differed somewhat relative to AMT respondents with respect to gender, age, education, and income.

The third data set was comprised of 138,998 user drug experience assessments collected from the AskAPatient online forum. The AskAPatient dataset was included due to its complementary nature to the AMT and Qualtrics datasets with respect to number of instances, dimensions, and response collection mechanism. Collectively, the testbed was comprised of a diverse array of datasets, tasks, and user content channels.

5 EVALUATION

5.1 Experiment Results – Benchmark Methods

In order to assess the performance of our PyNDA architecture, we conducted an extensive benchmark evaluation in comparison with 16 text classification techniques, presented in Table 2. The comparison methods belong to five categories: feature-based classifiers, CNNs, LSTMs, hybrid deep learning architectures, and multi-task deep learning methods. While the selected techniques are not an exhaustive list, they are representative of state-of-the-art approaches in each of the five categories. Included were well-established feature-based classifiers, such as the Multinomial Naïve Bayes [13], Logistic Regression [24], FRN [17], FastText, and Linear SVM [25]. Also selected were widely-used CNN architectures, such as CNNSent [26], CNNChar [27], VD-CNN [38], SWISSCHEESE [37], and SENSEI-LIF [14]. In order to further enrich the evaluation, we also built several custom CNN architectures, such as CNNWordRep which uses CNN to build word and representation embeddings before inserting these into the dense layers for final classification, and CNNCombine which uses CNN to build word and representation embeddings simultaneously. Prominent LSTM architectures were also selected, including the basic LSTM [30], LSTMWordRep [39], and LSTMCombine. The LSTMsThenCNNs [34] method was included as a hybrid deep learning architecture.

For feature-based classifiers, consistent with prior work, we adopted unigram features with tfidf weighting since these yielded the best performance. Additionally, for multinomial Naïve Bayes, we used Laplace smoothing, and we ran Logistic regression using the L2 penalty with LibLinear solver with 100 maximum iterations. For SVM, we adopted the L2 penalty with squared hinge loss function, with 1000 maximum iterations. For all the deep learning benchmarking models, we adopted the settings and parameter values from the original studies. Moreover, the text representations for all benchmarking deep learning models were also those used in the original studies.

The parameter settings for our proposed architecture will be provided upon requests.

Bifurcation was performed on each dataset to convert the continuous psychometric target class variables into binary classification variables. Consistent with prior studies, this was done by only using instances from the end quartiles as the low and high class labels, respectively.

Table 2 presents the PyNDA results along with the 16 benchmarking methods, averaged across the eleven datasets. PyNDA significantly outperformed the benchmarking methods across all evaluation metrics, including accuracy, precisions, recalls and receiver operating characteristic curve area-under-the-curve (ROC). The overall accuracy, F-measures, and ROC for PyNDA were at least 5 percent higher than the second best method. Among benchmarking methods, LSTM architectures were better than CNN, underscoring the importance of capturing long-term dependencies among texts for more effective psychometric classification. CNNChar yielded the worst results, suggesting that morphological patterns may not be critical indicators for psychometric-related texts. Instead, word or sentence level features may have more predictive power, as illustrated by the relatively higher performance for CNNSent and CNNWord. Given the recent effectiveness of CNN approaches in sentiment classification tasks (e.g., [14], [37]), the relative superiority of LSTM in our context reinforces previously stated notions of complexity of the nuanced psychometric dimensions examined in our study. Feature-based classifiers demonstrated reasonable performance, relative to alternative benchmarking methods.

TABLE 2
SUMMARY OF BENCHMARKING RESULTS

Category	Method	Accuracy	Precision+	Precision-	Recall+	Recall-	F-measure+	F-measure-	ROC
Feature	FastText [41]	73.9	73.4	74.4	74.9	72.9	74.2	73.6	73.9
	FRN [17]	74.9	73.0	75.0	74.3	74.8	73.7	74.9	74.9
	Linear SVM [25]	73.5	73.3	73.7	73.8	73.1	73.6	73.4	73.5
	Logistic Regression [24]	74.6	73.9	75.6	76.1	73.2	75.0	74.3	74.7
	Multinomial Naive Bayes [23]	73.4	72.9	74.8	75.7	71.2	74.3	72.9	73.5
CNN	CNNSent [26]	74.4	73.4	76.4	77.9	70.8	75.6	73.5	74.4
	CNNChar [27]	66.6	66.2	68.0	69.4	63.7	67.8	65.8	66.6
	CNNWordRep	73.6	73.4	74.4	74.7	72.6	74.0	73.5	73.6
	CNNCombine	73.2	72.2	74.7	75.2	71.1	73.7	72.8	73.2
	SENSEI-LIF [14]	73.4	73.2	74.1	74.4	72.3	73.8	73.2	73.4
	SWISSCHEESE [37]	74.1	73.6	73.9	76.3	71.8	74.9	72.9	74.1
	VeryDeepCNN [38]	58.9	58.1	60.4	63.2	54.2	60.5	57.1	58.7
LSTM	LSTM [30]	72.6	72.5	73.5	73.5	71.8	72.9	72.6	72.6
	LSTMCombine	74.8	74.8	75.1	75.5	74.1	75.2	74.6	74.8
	LSTMWordRep [39]	74.5	74.6	74.7	74.6	74.3	74.6	74.5	74.4
Hybrid	LSTMsThenCNNs [46]	75.1	74.1	76.8	77.6	72.5	75.8	74.5	75.1
PyNDA		**81.1**	**80.2**	**82.1**	**82.7**	**79.4**	**81.4**	**80.7**	**81.0**

5.2 Experiment Results – Ablation Analysis

In order to examine the additive impact of each component of the architecture, ablation analysis was performed. The top half of Table 3 shows the summary (averaged) results across all eleven tasks associated with our three test beds. Paired t-test results revealed that each additional component significantly enhanced performance over the prior ablation setting (all p-values < 0.05).

We also evaluated two alternative MTL setups. The first was a separate word embedding and separate and shared LSTMs, exactly as proposed in [40], concatenated with the rest of our architecture. The second was our MTL with only the final concatenation layer (i.e., no embedding LSTM-level weights). The results, depicted in the bottom of Table 3, show that the more holistic application of MTL in PyNDA, with inclusion of finer-grained component-level weights in L_{Task} coupled with training instance alignment, boosts accuracy by 1.5% to 4% over alternative setups.

TABLE 3
SUMMARY OF ABLATION ANALYSIS RESULTS

Ablation Setting	Acc	Prec+	Prec-	Rec+	Rec-
CharEmbeddingCNN	66.6	66.2	68.0	69.4	63.7
+ParallelRepsLSTM	74.8	74.8	75.1	75.5	74.1
+RepEmbedding	79.4	78.9	80.1	80.6	78.2
+DemEmbd&SEMEnc	80.3	79.5	81.3	82.0	78.6
+MultiTaskLearning	81.1	80.2	82.1	82.7	79.4
Alternative MultiTask Learning (MTL) Setups					
MTLSeparateWord	77.0	77.1	77.2	76.9	77.0
MTLNoComponents	79.6	79.3	80.0	80.5	78.7

6 CONCLUSION

In this paper, we propose a novel deep learning architecture, PyNDA, to extract both critical psychometric dimensions—including health literacy, health numeracy, trust in doctors and anxiety of seeing physicians—and drug experience assessments from user-generated texts. Our experiments on eleven tasks pertaining to three datasets show that PyNDA markedly outperforms traditional feature-based classifiers as well as state-of-the-art deep learning architectures. Given the lack of prior work focused on AI/machine-learning methods for deriving psychometrics from secondary data, the results represent a major contribution to the broader "knowledge and data engineering"field. Furthermore, PyNDA has profound practical implications: for example, it could be used to infer users' psychometric attitudes and beliefs, which drive key behaviors in various critical contexts such as health, cybersecurity, and e-commerce. Within the health domain, such models could be deployed via mobile apps to help infer patients' mental statuses related to chronic diseases in a timelier manner, thereby helping physicians conduct informed decision making and also allowing patients to better self-regulate their health statuses. In the future, we hope to extend our model to some of the other aforementioned application domains and also to deploy them in real-time synchronous chat contexts.

References

[1] D.E. Brown, A. Abbasi, and R.Y.K. Lau "Predictive Analytics: Predictive Modeling at the Micro Level," *IEEE Int. Syst.*, vol. 30, no. 3, pp. 6-8, 2015.

[2] A. Abbasi, R.Y.K Lau, and D.E. Brown, "Predicting Behavior," *IEEE Int. Syst.*, vol. 30, no. 3, pp. 35-43, 2015.

[3] H. Taylor, F. Henderson, A. Abbasi, and G. Clifford, "Cardiovascular Disease in African Americans: Innovative Community Engagement for Research Recruitment and Impact," *American Journal of Kidney Diseases*, 72(5)(Supp 1), 2018, S43-S46,

[4] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect Analysis of Web Forums and Blogs using Correlation Ensembles," *IEEE Trans. Knowledge and Data Engr*, 20(9), 2008, pp. 1168-1180.

[5] J. Zhou, et al., "Clustered multi-task learning via alternating structure optimization*," Proc. Advances in NIPS*, 2011, pp. 702-710.

[6] C. Che, et al., "An RNN Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson's Disease," *Proc. SIAM Int. Conf. on Data Mining*, 2017, pp. 198-206.

[7] N.D. Berkman, S. Sheridan, K. Donahue, D. Halpern, K. Crotty, "Low Health Literacy and Health Outcomes: An Updated Systematic Review", *Ann. of Internal Medicine*, vol. 155, pp. 97-107.

[8] E. Dugan, F. Trachtenberg, et al. "Development of Abbreviated Measures to Assess Patient Trust in a Physician and the Medical Profession," *BMC Health Services Res.*, vol. 5, no. 64, 2005.

[9] M.M. Schapira, C.M. Walker, et al. "Development and Validation of the Numeracy Understanding in Medicine Instrument Short Form," *J. of Health Commun.*, vol. 19, no. 2, pp. 240-253, 2014.

[10] D. Gefen and K. Larsen "Controlling for Lexical Closeness in Survey Research," *J. Assoc. for Inform. Syst.*, vol. 18, no. 10, pp. 727-757, 2017.

[11] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Syst.*, vol. 28, no. 2, pp. 15-21, 2015.

[12] Y. Zhang, Y. Dang, and H. Chen, "Gender Classification for Web Forums," *IEEE Trans. Syst., Man, and Cybernetics-Part Ams and Humans*, vol. 41, no. 4, pp. 668-677, 2011.

[13] J. Dressel and H. Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Sci. Advances*, vol. 4, no. 1, 2018.

[14] L. Xu, C. Jiang, Y. Ren, and H.H. Chen, "Microblog Dimensionality Reduction—A Deep Learning Approach," *IEEE Trans. Knowledge and Data Engr.*, vol. 28, no. 7, pp. 1779-1789, 2016.

[15] Z. Yu, H. Wang, X. Lin, and M. Wang, "Understanding Short Texts through Semantic Enrichment and Hashing," *IEEE Trans. Knowledge Data Engr.*, vol. 28, no. 2, pp. 566-579, 2016.

[16] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation," *ACM Trans. Manage. Inform. Syst.*, 9(2), 2018, no. 5.

[17] M. Rouviern and B. Favre, "SENSEI-LIF at SemEval-2016 Task 4: Polarity Embedding Fusion for Robust Sentiment Analysis," *Proc. of the 10th Int. Workshop on Semantic Evaluation*, pp. 202-208, 2016.

[18] A. Abbasi, S. L. France, Z. Zhang, and H. Chen, "Selecting Attributes for Sentiment Classification using Feature Relation Networks," *IEEE Trans. Knowledge and Data Engr.*, vol. 23, no. 3, pp. 447-462, 2011.

[19] E. Riloff, et al., "Feature Subsumption for Opinion Analysis," *Proc. of the 2006 ACL Conf. on Empirical Methods in Natural Language Processing*, pp. 440-448, 2006.

[20] R.H. Osborne, et al., "The Grounded Psychometric Development and Initial Validation of the Health Literacy Questionnaire (HLQ)," *BMC public health*, vol. 13, no. 1, 2013, pp. 658.

[21] P.J. Ciampa, et al., "Patient Numeracy, Perceptions of Provider Communication, and Cancer Screening," *J. Health Comm.*, vol. 15, no. sup3, 2010, pp. 157-168.

[22] E. Dugan, et al., "Development of Abbreviated Measures to Assess Patient Trust in a Physician and the Medical Profession," *BMC Health Services Res.*, vol. 5, no. 1, 2005, pp. 64.

[23] C.D. Spielberger, "*State-trait Anxiety Inventory: a Comprehensive Bibliography*, Consulting Psychologists Press, 1989.

[24] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *Proc. AAAI-98 workshop on learning for text categorization*, Citeseer, 1998, pp. 41-48.

[25] A. Genkin, et al., "Large-scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, vol. 49, 3, 2007, pp. 291-304.

[26] B. Pang, et al., "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," *Proc. ACL Conf. Empirical Methods in Natural Language Processing*, 2002, pp. 79-86.

[27] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746-1751, 2014.

[28] X. Zhang, et al., "Character-level Convolutional Networks for Text Classification," *Proc. Advances in Neural Information Processing Syst.*, 2015, pp. 649-657.

[29] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, 2001, pp. 5-32.

[30] D. Kaplan, *Structural Equation Modeling: Foundations and Extensions*, Sage Publications, pp 79-80, 2008.

[31] J. Li, et al., "When are Tree Structures Necessary for Deep Learning of Representations?,"*Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing*, pp. 2304-2314. 2015.

[32] J.L. Elman, "Finding Structure in Time," *Cognitive Sci.*, vol. 14, no. 2, pp. 179-211, 1990.

[33] Y. LeCun, et al., "Gradient-based Learning Applied to Document Recognition," *Proc. IEEE*, vol. 86, no. 11, 1998, pp. 2278-2324.

[34] C. Zhou, et al., "A C-LSTM Neural Network for Text Classification," *arXiv:1511.08630*, 2015.

[35] R. Collobert and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," *Proc. 25th Int. conf. on Mach. Learning*, ACM, 2008, pp. 160-167.

[36] I. Goodfellow, et al., "Generative Adversarial Nets," *Proc. Advances Neural Information Processing Systems*, 2014, pp. 2672-2680.

[37] C.D.V. Hoang, et al., "Incorporating Side Information into Recurrent Neural Network Language Models," *Proc. 15th Conf. NAACL Human Language Technologies*, 2016, pp. 1250-1255.

[38] J. Deriu, et al., "Swisscheese at SEMEVAL-2016 Task 4: Sentiment Classification Using an Ensemble of Convolutional Neural Networks with Distant Supervision," *Proc. 10th Intl. Workshop Semantic Evaluation*, 2016, pp. 1124-1128.

[39] A. Conneau, et al., "Very Deep Convolutional Networks for Text Classification," *Proc. 15th Conf. European Chapter ACL: Volume 1, Long Papers*, 2017, pp. 1107-1116.

[40] P. Wang, et al., "A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding," *arXiv:1511.00215*, 2015.

[41] P. Liu, et al., "Adversarial Multi-task Learning for Text classification," *Proc. of the 55th Annual Meeting of the ACL*, vol. 1, pp. 1-10, 2017.

[42] A. Joulin, et al., "Bag of Tricks for Efficient Text Classification," *Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics*, vol. 2, pp. 427-431, 2017.

[43] J. Li, K. Larsen, and A. Abbasi, "TheoryOn: A Design Framework and System for Unlocking Behavioral Knowledge through Ontology Learning," *MIS Quarterly*, forthcoming.

[44] F. Ahmad, A. Abbasi, J. Li, D. Dobolyi, R. Netemeyer, G. Clifford, and H. Chen, "A Deep Learning Architecture for Psychometric Natural Language Processing," ACM Transactions on Information Systems, forthcoming.

[45] A. Abbasi, J. Li, D. Adjeroh, M. Abate, and W. Zheng "Don't Mention It? Analyzing User-generated Content Signals for Early Adverse Event Warnings," *Information Systems Research*, 30(3), 2019, pp. 1007-1028.

[46] R. Netemeyer, D. Dobolyi, A. Abbasi, G. Clifford, and H. Taylor, "Health Literacy, Health Numeracy and Trust in Doctor: Effects on Key Consumer Health Outcomes," *Journal of Consumer Affairs*, forthcoming.

[47] A. Abbasi, J. Li, G. D. Clifford, and H. A. Taylor, "Make 'Fairness By Design' Part of Machine Learning," *Harvard Business Review*, August 5, 2018, digital article: https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning

[48] B. Kitchens, D. Dobolyi, J. Li, and A. Abbasi, "Advanced Customer Analytics: Strategic Value through Integration of Relationship-Oriented Big Data," *Journal of Management Information Systems*, 35(2), 2018, pp. 540-574.

[49] J. Li, K. Larsen, and A. Abbasi, "Unlocking our Behavioral Knowledge Inheritance through Ontology Learning: A Design Framework, an Instantiation, and a Randomized Experiment," *In the INFORMS Workshop on Data Science*, Houston, TX, Oct. 21, 2017.

[50] J. Li, K. Larsen, and A. Abbasi, "Unlocking Knowledge Inheritance of Behavioral Research through Ontology Learning: An Ontology-Based Search Engine," *In the 26th Workshop on Information Technologies and Systems (WITS)*, Dublin, Dec 15-16, 2016.

[51] A. Abbasi, Y. Zhou, S., Deng, and P. Zhang, "Text Analytics to Support Sense-making in Social Media: A Language-Action Perspective," *MIS Quarterly*, 42(2), 2018, pp. 427-464.

[52] J. Li, A. Abbasi, F., Ahmad, and H. Chen, "Deep Learning for Psychometric NLP," *In the 28th Workshop on Information Technologies and Systems (WITS)*, Dec 15-16, 2018.

[53] J. Li, A. Abbasi, A. Cheema, and L. Abraham, "Path to Purpose? Impact of Online Purchases' Hedonic and Utilitarian Characteristics on the Customer Journey," *In the 26th Workshop on Information Technologies and Systems (WITS)*, Dublin, Dec 15-16, 2016.