Interpretable Representation Learning for Personality Detection

Amirmohammad Kazemeini, Sudipta Singha Roy, Robert E. Mercer
Department of Computer Science
The University of Western Ontario, Canada
{akazemei,ssinghar,mercer}@uwo.ca

Erik Cambria
School of Computer Science and Engineering
Nanyang Technological University, Singapore
cambria@ntu.edu.sg

***Abstract*—Automatic personality detection has gained increasing interest recently. Several models have been introduced to perform this task. The weakness of these models is their inability to interpret their results. Even if the model shows excellent performance over test data, it can sometimes fail in real-life tasks since it may incorrectly interpret a statement. To investigate this issue, we evaluate two approaches. In the first approach, we generate sentence embeddings by training a siamese Bi-LSTM with max-pooling on the psychological statement pairs to compute the semantic similarities between them. In the second approach, we evaluate state-of-the-art pretrained language models to see whether their output representations can distinguish personality types. Both of these approaches outperform state-of-the-art models for this task with less computational overhead. We conclude by discussing the implications of this work for both computational modeling and psychological science.**

***Index Terms*—Personality detection, Interpretability, BERT**

I. Introduction

AI has the potential to assist health experts in dealing with the increasing rate of mental health issues and disorders. This increasing trend has been the subject of recent investigations such as the recent trends in mental ill health and health-related behaviors in two cohorts of UK adolescents that show depressive symptoms and self-harm were higher in 2015 compared with 2005 [1]. How social media impacts mental health has also been studied [2]. This increasing rate of mental issues has accelerated due to the COVID-19 pandemic. According to a Kaiser Family Foundation poll, people have become more socially isolated and stressed. Nearly half of Americans report the coronavirus crisis is harming their mental health [3], [4].

According to a 2020 Harris Poll, between 46% and 51% of US adults were using social media more since the outbreak began [5]. Increased social media use means more digital footprints, and since people's personality and private traits can be identified based on them [6], this pandemic challenge can be turned into an advantage to provide more support for people based on their needs. A WHO survey showed that COVID-19 further burdened the already limited mental health services in many countries [7]. Since mental health service resources are limited and mental health issues have increased, the increase in social media use provides an opportunity for AI researchers to utilize the produced digital footprints to help diagnose people's mental health issues.

Personality traits are defined as the set of relatively stable characteristics which describe our feelings and behavior. These traits play important roles in individuals' futures and life outcomes [8], [9]. Among the various personality tests, the Big-Five, which is also called OCEAN, is known to be the most reliable test for assessing people's personality [10]. The OCEAN test describes personality in five measures: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Previous work has investigated the relationship between personality and mental disorders. Studies have shown that neuroticism plays a vital role in depressive and anxiety disorders [11]. Regarding the other traits, resilience demonstrates a strong inverse relationship with neuroticism and strong positive relationships with extraversion and conscientiousness and a small but statistically significant positive relationship with openness [12]. Hence, understanding a person's personality can provide a better insight for detecting mental illnesses. In addition to psychological motivation, personality traits are also useful in recommendation systems [13]–[15], product and service personalization [16], [17], job screenings [18], social network analysis [19], and sentiment analysis [20].

In this work, we address the following two questions: Does the embedding, which is used for current state-of-the-art model, capture psychological information? If not, how can it be improved? In order to answer these questions, we first introduce an approach for evaluating embeddings in personality detection. Following that, with metric learning in mind [21], we apply two different approaches using two siamese architectures for generating the embeddings from the psychological statements. The first approach produces sentence embeddings by means of computing semantic similarities between psychological statements representing different traits. In the second approach, different variants of another siamese sentence encoder, Sentence-BERT, for producing sentence embeddings for classifying psychological traits are investigated. Both of these approaches surpass the previous state-of-the-art models used in this task with the BFI statement data [22]–[24]. The second approach outperforms the previous state-of-the-art models with the Essays dataset [25] and the Kaggle personality dataset [26]. Extensive experiments with the Essays dataset and the BFI statements are performed and discussed. These experiments have focussed on these two datasets since the MBTI test (used in the Kaggle personality dataset) has been questioned for

2375-9259/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDMW53433.2021.00026

its comprehensiveness, dependability, and lack of independent categories [27], whereas the OCEAN personality test (the Essays dataset) is considered as more reliable. These approaches[1] not only outperform the previous state-of-the-art model but also reduce the computational overhead.

II. Related Work

There are a variety of personality tests that are based on psychological discoveries [28]. The most accepted one in the field of psychology is the Big Five model, also called OCEAN [10]. This personality test is the one focussed on in this paper. OCEAN assesses five dimensions of personality (Openness to Experience, Conscientiousness, Agreeableness, Extraversion, and Neuroticism or when positively keyed, emotional stability). One other commonly used personality model, which is used in a comparison below, is Myers-Briggs, also known as MBTI [29]. MBTI categorizes personalities into 16 types; each one can be described as a combination of 4 binary categories (Extroversion/Introversion, Sensing/Intuition, Thinking/Feeling, Judging/Perceiving). Since the MBTI test has been questioned for its comprehensiveness, reliablity, and lack of independent categories, the OCEAN personality test is chosen as the main focus of this paper.

Given the limited mental health service resources, there is a strong need for an automated assistant tool. AI models have proven to be good candidates as they perform more accurately than humans in personality judgment [30]. Some models used psycholinguistic features to identify personality [31]. In the field of deep learning-based automatic personality detection, the hierarchical CNN model [32] and multitask learning [33] have attracted a lot of attention. A full comparison between previous proposed models is given in [34] and perspectives are analyzed in [35]. Although the deep models are improving the accuracy in this field and their approaches have built the foundations of our current work, they suffer from some issues that prevent them from serving as well as they ought to. For example, the results might be based on the studied socio-cultural group. Lewis [36] has analyzed this diversity and has shown that the results can vary depending on the observed cohort. In addition, due to the delicate nature of mental health tasks, trust is an important criterion that these black-box models cannot satisfy without using a post-hoc explainability approach [37].

Current NLP models that understand human language are mostly proposed by large companies such as Facebook and Google, enabled by their high-spec infrastructure to create their high accuracy predictors [38]–[40]. Although they are not runnable on regular computers, their pre-trained versions can be used in personality detection with a small amount of fine-tuning to be adapted to this task [41], [42]. Considering that there is usually a trade-off between accuracy and simplicity, the task to obtain an optimal, yet simple model is non-trivial. Only a few papers, such as [43] (BB-SVM), have proposed high accuracy models in this field without sacrificing simplicity. BB-SVM also introduced a BERT-based personality model that can be used for longer sequences as well. However, even though this model is able to be run on ordinary computers, its interpretability, especially the justification for the choice of the pre-trained model, has yet to be addressed. First, as well as the existing trade-off between complexity and accuracy, a trade-off also exists between performance and transparency (i.e., explainability of the outcomes). The higher performing models tend to be more opaque [44]. As the model becomes more opaque, the need for explainability increases. To alleviate this problem, post-hoc explainability is used. This type of explainability is divided into model-agnostic approaches, which can be used for any model, and model-specific ones. A full comparison of explainable AI methods is given in [45]. Also, contemporary models learn from examples in specific datasets. This issue challenges the model when it faces new examples that are not the same as the previously observed ones since current models are not using experts' knowledge. So, even though the current models can do their best for their specific dataset, they cannot incorporate the socio-cultural diversity among groups of people, which results in the different ways they articulate their thoughts [36].

With the emergence of accurate AI models, theorists and researchers make normative claims based on the models' results [46]. Some of the previous experience has also shown how these models can be exploited for detrimental goals [14], [47]. Hence, by making the AI models more interpretable, more descriptive facts can be obtained based on their results. Ethical concerns can be slightly alleviated because of the insight which the model provides. [41] is one of the few works that address both improving personality detection accuracy using deep learning models and providing understandable insight using post-hoc explainablity approaches. This work is used as the baseline for the current paper.

III. Methodology

This section discusses the interpretable sentence representation generation approaches using the siamese architectures, the dataset we use for training the model, and the datasets used for evaluating the performance of the models. The sentence representation is generated by means of computing the semantic similarities between psychological statements. The reason behind choosing this approach is to preserve enriched semantics in the vector representations. Finally, the approach to interpret the output of the model is discussed along with the evaluation of the model. The interpretability of our approach is evaluated using the feature relevance and visual explanation methods of the post-hoc explainability category [45], by computing the cosine similarity between the input and baseline sentences and using PCA visualization, respectively.

A. Datasets

In this work, we used the Essays dataset [25], which consists of 2468 essays written by students and annotated according to the Big Five personality, and the Kaggle dataset, a collection of 8675 records collected from the PersonalityCafe forum [26].

[1]Code: https://github.com/amirmohammadkz/interpretable_personality

B. Evaluating the Embeddings

In order to evaluate the pretrained BERT-base model for meaningful personality representations, we have used a simplified version of the Big Five Inventory (BFI) [22]–[24]. BFI is a self-report questionnaire that consists of 44 short phrases. Participants rate each of these statements based on their situation. Each statement focuses on assessing one of the five traits. We have simplified this version to make it easier for language models to extract meaningful representations from them. For example, the statement "I am someone who is talkative", which assesses the extraversion rate of a person, is converted to "I am talkative". In addition, to increase the dataset size, we have also added the adapted version of BFI [48], [49] to the original one. The final simplified statement set consists of 85 sentences, 44 of which belong to the original BFI statements and the rest are obtained from the adapted version. We then use the pretrained version of BERT-base to extract the representations of the tokens.

We have followed the best representation of [41] which is averaging the output of the second to last layer to get the final representation of each statement. Next, we transform the embeddings using a PCA [50] with 2 principal components. The result of the PCA is illustrated in Fig. 1. The B-points are clustered in the upper half of the bottom right quadrant, whereas the 0- and 1-points are almost all in the left or upper quadrants. The representations of the baseline sentences are very close to each other and the distance between them and the corresponding trait statements are much larger. Hence, we can conclude that even when [41] gets high accuracy using these representations, it will not be generalizable since the extracted embeddings do not manifest the related personalities. Considering that this current state-of-the-art representation uses a rich corpus and state-of-the-art language models, we can infer that older ones probably also suffer from this issue. Furthermore, even if the baseline representations obtained

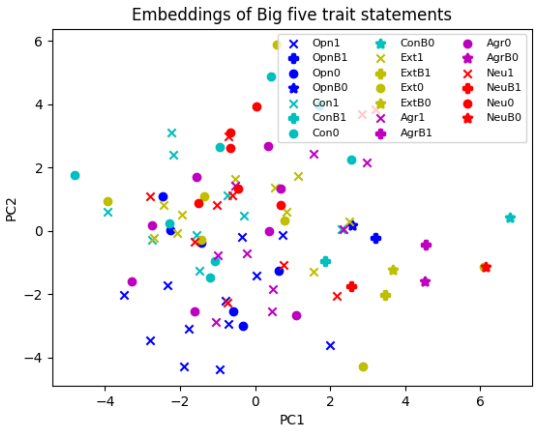

Fig. 1. Visualization of the personality statements after applying PCA on the average of the output of layer 11 of Bert-base [41]. 1 and 0 mean "High" and "Low" rate of a specific trait, respectively, and "B" is for baseline sentences.

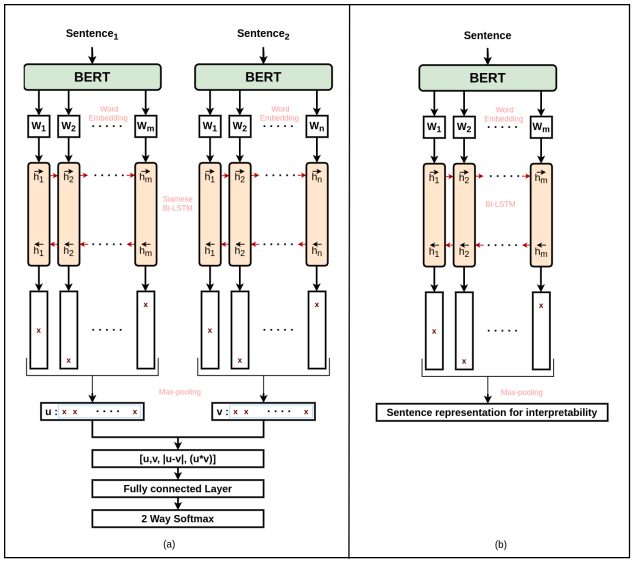

Fig. 2. Architecture of the model with siamese Bi-LSTM and max-pooling for the interpretable tool for personality detection. (a) The training of the model, (b) After training, the Bi-LSTM followed by the max-pooling layer act as the sentence encoder.

from the previous methods maintain sufficient distance, their classification performance is worse compared to [41] which is also not acceptable. This motivates our investigating a model which cannot only improve the classification performance but also enhance explainability.

C. Interpretable Representation for Personality Detection

This paper investigates two different approaches for producing vector representations from psychological statements. The core idea behind both approaches is to use the extracted embeddings from the baseline sentences and BFI statements in order to evaluate the performance of the model. The output embedding can be explainable using this comparison.

Both of these approaches use siamese architectures using deep learning models. The first approach utilizes siamese Bi-LSTM with max-pooling over time of the output vectors. This model is trained on the simplified BFI statement pairs for computing the similarity between them. The second approach evaluates the Sentence-BERT variants [51]. The reason behind choosing the siamese models here is that we try to detect the personality traits not by applying direct classification approaches but rather by preserving the semantics of the statements where statements reflecting similar traits remain close to each other in the embedding space. This objective is achieved by leveraging psychological datasets (the BFI statements and the baseline sentences).

1) Bi-LSTM with Max-pooling: To extract the feature vectors of both the BFI statements and the baseline sentences, we have used the siamese architecture of Bi-LSTM over the BERT word embeddings from layer 11 of BERT-base. The architecture is inspired by the InferSent model [52]. The basic idea of this model is to generate a sentence embedding

by means of computing the semantic similarity between two sentences. This semantics attempts to preserve the personality trait from the BFI statement. For the word embeddings, we have chosen the output of layer 11 of the pre-trained BERT-base. For any given sentence pair, word embeddings are fed to two identical Bi-LSTMs. These Bi-LSTMs share the same parameters and weights. For a sequence of N words, Bi-LSTM produces a set of N vectors. The final hidden state representation for each time step is generated by concatenating the hidden representation of the forward ($\overrightarrow{h_i}$) and backward LSTMs ($\overleftarrow{h_i}$) [53]. For each time step, max-pooling is applied over these concatenated hidden representations ($[\overrightarrow{h_i}, \overleftarrow{h_i}]$) to generate an intermediate sentence representation. In the next step, three operations, concatenation, point-wise difference and point-wise multiplication, are performed on the representations obtained for both of the sentences from the sentence pair. Finally, the outcome of these three matching operations are concatenated and fed to a feed-forward neural network for classification like [52]. Suppose, u and v are the intermediate representations for the sentences after max-pooling. Then $[u, v, |u-v|, (u*v)]$ would be the final feature representation to be fed to the following classifier. The classifier outputs either 0 or 1 where 1 indicates the sentences offer semantically similar traits and 0 otherwise. Fig. 2 portrays the overall architecture of the model. After the training is done, the Bi-LSTM together with the max-pooling layer acts as the encoder for generating the sentence representation. This representation is a 768 dimensional vector.

2) Sentence-BERT: Sentence-BERT [51] is a refinement of the pretrained BERT using siamese and triplet structures. It can derive sentence representations preserving the semantics of the sentences. Unlike BERT, which outputs rich token embeddings and [CLS] with poor semantics for the sentence, Sentence-BERT produces semantically richer sentence embeddings. It is trained on the sentence pairs from the SNLI dataset [54] and multi-genre NLI dataset [55]. It has been shown that sentence embedding models trained on natural language inference datasets have better semantic preserving abilities [52]. For this reason, Sentence-BERT outputs semantically richer sentence embeddings. Sentence-BERT incorporates a mean-pooling operation over the output of each BERT embedding to generate two sentence embeddings for the sentence pair. Then two matching operations, concatenation and point-wise difference, are performed on them. Finally, this feature is fed to the softmax classifier. After the fine tuning is complete, the fine-tuned BERT with the mean-pooling act as the sentence encoder. Using this pretrained Sentence-BERT is then a straightforward approach. After being given a sentence, it directly outputs the corresponding 768 dimensional vector sentence embedding. The architecture of Sentence-BERT is shown in Fig. 3. We have conducted experiments on the Essays, the BFI statements, and the Kaggle datasets using different variants of Sentence-BERT [56]. In all cases the overall architecture remains the same, only the BERT encoder is varied. Some prominent variants are RoBERTa [40] and MPNet [57].

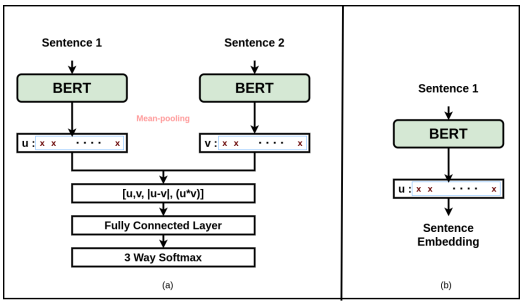

Fig. 3. Architecture of Sentence-BERT. (a) Training of the model on the natural language inference datasets. (b) Sentence encoder.

IV. Experiments

To analyze the effectiveness of our siamese Bi-LSTM model, for each personality trait t, we create all possible corresponding BFI statement pairs together with the appropriate label, $(s_i, s_j, l_{i,j})$, where $l_{i,j}$ is 1 if the statements s_i and s_j have the same label and 0 if s_i and s_j have different labels. Then, we feed the statement pairs as inputs to the model and use $l_{i,j}$ as the label which the model tries to predict. Applying this approach over the BFI statements, the data set has 681 sentence pairs. Among these, 600 samples are used for training and the remaining 81 are used for validation. This small dataset was sufficient for training the siamese LSTM model with some good training and validation accuracies. While testing this model on the BFI statements, it achieved a better result compared to the previous models [41]. This comparison is performed using the *PredLabel* and *SimScore* metrics. In addition, the finetuned embedding are also assessed by replacing the embedding part of the model in [41] for classifying the Kaggle and Essays datasets. However, the model trained on this data did not achieve state of the art accuracies as the training data was comparably small.

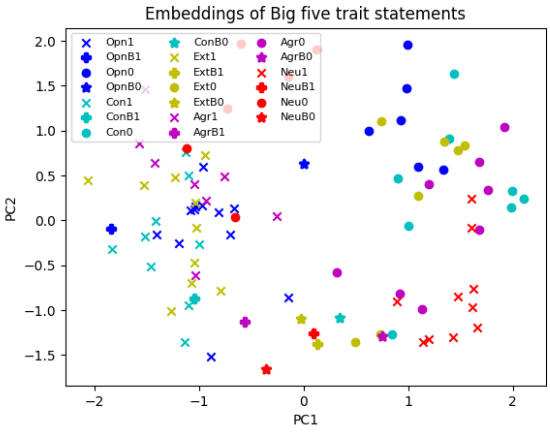

Fig. 4. Visualization of the personality statements after applying PCA on the feature vectors of Bi-LSTM and max-pooling. 1 and 0 mean "High" and "Low" rate of a specific trait, respectively, and "B" is for baseline sentences.

We have trained the siamese Bi-LSTM model for only 25 epochs where the best result was found at the 21st epoch. While training, the batch size was set to 10 with 10% dropout. Standard gradient descent was used for optimization with a learning rate $1e^{-5}$. The forward and backward LSTMs' hidden representations are 384 dimensional vectors. After the training phase, we use the feature vectors extracted from the Bi-LSTM for evaluation as we did in Section III-B for the BFI statements. After extracting the feature vectors of both the BFI statements and the baseline sentences, for each statement that belongs to trait t we assign a similarity score and prediction label based on the closeness to the corresponding baseline sentences as follows:

$$\forall s_i \in S_t : \textit{SimScore}(s_i) = (-1)^{l_i - 1} C(s_i, b_{t,1}) + (-1)^{l_i} C(s_i, b_{t,0})$$

and

$$\textit{PredLabel}(s_i) = \begin{cases} 1, & \text{if } C(s_i, b_{t,1}) > C(s_i, b_{t,0}) \\ 0, & \text{otherwise} \end{cases}$$

where l_i is the label of s_i, C is cosine similarity, and $b_{t,0}$, $b_{t,1}$ are the baseline feature vectors of trait t. To report the result of a specific model, we use accuracy for the *PredLabel*s and the average of the *SimScore*s. For the Sentence-BERT models, the BFI statements and baseline statements are fed to the pretrained encoders and then the accuracy of the *PredLabel*s and the average of the *SimScore*s are computed. While testing, we aggregated both the simplified and non-simplified versions of the BFI statements to generate a more generalized model. The embeddings of the BFI and the baseline statements are extracted from the encoder portion of the siamese Bi-LSTM as previously described and finally, *PredLabel*s and *SimScore*s are measured. In the case of experimenting with the Essays dataset, no further training is performed. The statements are fed to the models (both the Bi-LSTM with max-pooling and the Sentence-BERTs). Then they are tested against the baseline statements to compute the performance metrics. The Kaggle dataset is tested with the Sentence-BERTs only.

V. Results

The accuracies of the *PredLabels* are shown in Table I, and the *SimScore*s for the BFI statements, in Table II. For three traits, Bi-LSTM with max-pooling outperforms the CLS and average methods of BERT which were used in [41]'s state-of-the-art model for this task and outperforms on the average result as well. For each of the personality traits, the 0- and 1-statements form distinguishable and well-separated clusters except for the Neuroticism and Extroversion baseline sentences, which are so close to each other. The PCA result is illustrated in Fig. 4. The evaluation also tries to identify whether the model is able to assign the correct binary trait label to the statements. For Openness, Conscientiousness, and Agreeableness, as it is shown in Fig. 4, the model can almost completely understand which statement belongs to which baseline trait.

TABLE I
COMPARISON OF ACCURACIES OF *PredLabel*S OF DIFFERENT REPRESENTATIONS.

Model	O	C	E	A	N	Average
BERT (average) [41]	61.11	52.94	41.18	64.71	56.25	55.24
BERT (CLS)	33.33	58.82	41.18	47.06	62.5	48.58
Bi-LSTM with max-pooling	94.44	100.00	32.35	100.00	53.13	75.98
avg_word_embeddings_glove.6B.300d	33.33	58.82	70.59	76.47	43.75	56.59
avg_word_embeddings_glove.840B.300d	33.33	64.71	88.24	70.59	62.50	63.87
avg_word_embeddings_komninos	33.33	70.59	76.47	70.59	75.00	65.20
avg_word_embeddings_levy_dependency	33.33	41.18	47.06	64.71	62.50	49.76
nli-bert-base	66.67	76.47	70.59	88.24	100.00	80.39
nli-bert-base-cls-pooling	77.78	76.47	70.59	88.24	93.75	81.36
nli-bert-base-max-pooling	77.78	88.24	70.59	88.24	93.75	83.72
nli-bert-large	94.44	94.12	100.00	88.24	93.75	94.11
nli-bert-large-cls-pooling	88.89	88.24	100.00	88.24	100.00	93.07
nli-bert-large-max-pooling	88.89	82.35	100.00	88.24	100.00	91.90
nli-distilbert-base	72.22	88.24	17.65	88.24	93.75	72.02
nli-distilbert-base-max-pooling	77.78	82.35	11.77	88.24	87.50	69.53
nli-distilroberta-base-v2	72.22	94.12	70.59	88.24	100.00	85.03
nli-mpnet-base-v2	100.00	88.24	94.12	94.12	93.75	94.04
nli-roberta-base	94.44	82.35	100.00	88.24	93.75	91.76
nli-roberta-base-v2	83.33	94.12	100.00	88.24	100.00	93.14
nli-roberta-large	100.00	100.00	100.00	88.24	100.00	**97.65**
paraphrase-distilroberta-base-v1	33.33	70.59	47.06	70.59	87.50	61.81
paraphrase-xlm-r-multilingual-v1	83.33	70.59	47.06	76.47	93.75	74.24
stsb-bert-base	72.22	76.47	76.47	76.47	87.50	77.83
stsb-bert-large	88.89	88.24	100.00	82.35	68.75	85.65
stsb-distilbert-base	72.22	88.24	29.41	82.35	93.75	73.19
stsb-distilroberta-base-v2	72.22	82.35	70.59	82.35	100.00	81.50
stsb-mpnet-base-v2	94.44	94.12	94.12	100.00	93.75	95.29
stsb-roberta-base	100.00	70.59	76.47	82.35	100.00	85.88
stsb-roberta-base-v2	88.89	70.59	88.24	88.24	100.00	87.19
stsb-roberta-large	100.00	94.12	76.47	88.24	100.00	91.77

TABLE II
COMPARISON OF *SimScore*S OF DIFFERENT REPRESENTATIONS.

Model	O	C	E	A	N	Average
BERT (average) [41]	0.011	0.007	-0.003	0.026	0.002	0.009
BERT (CLS)	0.001	0.001	-0.011	0.012	0.009	0.002
Bi-LSTM with max-pooling	0.082	0.064	-0.01565	0.079	0.008	0.044
avg_word_embeddings_glove.6B.300d	0.000	0.039	0.038	0.066	0.011	0.031
avg_word_embeddings_glove.840B.300d	0.000	0.036	0.040	0.082	0.077	0.047
avg_word_embeddings_komninos	0.000	0.036	0.031	0.039	0.077	0.036
avg_word_embeddings_levy_dependency	0.000	-0.007	0.002	0.020	0.075	0.018
nli-bert-base	0.124	0.148	0.073	0.253	0.321	0.184
nli-bert-base-cls-pooling	0.145	0.134	0.063	0.277	0.330	0.190
nli-bert-base-max-pooling	0.116	0.141	0.035	0.187	0.224	0.141
nli-bert-large	0.231	0.211	0.160	0.270	0.211	0.217
nli-bert-large-cls-pooling	0.224	0.166	0.159	0.281	0.304	0.227
nli-bert-large-max-pooling	0.163	0.169	0.246	0.283	0.264	0.225
nli-distilbert-base	0.068	0.149	-0.088	0.194	0.224	0.109
nli-distilbert-base-max-pooling	0.088	0.147	-0.082	0.162	0.166	0.096
nli-distilroberta-base-v2	0.037	0.119	0.046	0.180	0.181	0.112
nli-mpnet-base-v2	0.148	0.086	0.209	0.253	0.223	0.184
nli-roberta-base	0.194	0.158	0.142	0.228	0.356	0.215
nli-roberta-base-v2	0.160	0.117	0.138	0.206	0.226	0.169
nli-roberta-large	0.248	0.278	0.245	0.274	0.415	**0.292**
paraphrase-distilroberta-base-v1	0.020	0.025	0.002	0.060	0.080	0.037
paraphrase-xlm-r-multilingual-v1	0.032	0.030	-0.004	0.074	0.117	0.050
stsb-bert-base	0.158	0.129	0.150	0.200	0.212	0.170
stsb-bert-large	0.251	0.174	0.145	0.261	0.140	0.194
stsb-distilbert-base	0.119	0.163	-0.041	0.221	0.272	0.147
stsb-distilroberta-base-v2	0.045	0.131	0.039	0.196	0.227	0.128
stsb-mpnet-base-v2	0.174	0.081	0.206	0.191	0.179	0.166
stsb-roberta-base	0.259	0.095	0.152	0.305	0.352	0.233
stsb-roberta-base-v2	0.107	0.097	0.122	0.190	0.243	0.152
stsb-roberta-large	0.218	0.262	0.077	0.226	0.315	0.219

Regarding Neuroticism, although the *SimScore* is better than both the CLS and the average methods, the classification metric was not satisfactory. Extraversion also seems to be the most difficult trait to be identified by baseline sentences. Although the statements are separated, the embeddings of "I am extrovert" and "I am introvert" are still too close, resulting in the poor result. We believe this issue happens because of the dataset which is used for training BERT. Overall, since we have not used the baseline sentences in any phase of the training process, and they are used only in the evaluation, we believe that Bi-LSTM with max-pooling has used the general language model knowledge enriched with knowledge from the psychological statements to distinguish between traits.

TABLE III
ACCURACY OF BI-LSTM WITH MAX-POOLING AND SENTENCE BERT MODELS ON ESSAYS AND KAGGLE DATASETS.

MODEL	Essays						Kaggle MBTI				
	O	C	E	A	N	Average	I/E	N/S	T/F	P/J	Average
Majority Baseline	51.5	50.8	51.7	53.1	50.0	51.4	77.0	85.3	54.1	60.4	69.2
BERT-base [41]	64.6	59.2	60.0	58.8	60.5	60.6	78.3	86.4	74.4	64.4	75.9
BERT-large [41]	63.4	58.9	59.2	58.3	58.9	59.7	78.8	86.3	76.1	67.2	77.1
Bi-LSTM max-pooling_combined	61.7	54.6	55.0	56.7	55.9	56.8	-	-	-	-	-
average_word_embeddings_glove.6B.300d	63.2	58.5	56.3	57.2	58.5	58.7	77.2	86.5	76.9	66.2	76.7
average_word_embeddings_glove.840B.300d	63.0	58.0	57.2	57.5	57.7	58.7	78.6	87.1	**79.6**	68.6	78.5
average_word_embeddings_komninos	62.5	57.9	55.3	56.6	58.5	58.1	77.0	86.2	74.3	63.0	75.1
average_word_embeddings_levy_dependency	61.4	55.9	54.0	53.3	56.7	56.3	77.0	86.2	70.2	60.5	73.5
nli-bert-base	64.0	**60.0**	58.7	58.2	60.4	60.2	77.6	86.4	70.8	62.5	74.3
nli-bert-base-cls-pooling	63.8	59.7	57.7	59.1	60.1	60.1	77.6	86.3	71.1	62.2	74.3
nli-bert-base-max-pooling	63.0	58.0	56.7	57.4	58.4	58.7	77.5	86.2	69.7	61.8	73.8
nli-bert-large	63.5	59.8	57.1	58.7	60.8	60.0	77.6	86.3	71.2	62.2	74.3
nli-bert-large-cls-pooling	63.6	59.2	57.9	58.7	60.1	59.9	77.5	86.3	71.3	62.7	74.4
nli-bert-large-max-pooling	63.0	58.1	58.3	58.5	59.1	59.4	77.5	86.2	70.8	61.9	74.1
nli-distilbert-base	62.5	58.8	58.5	57.8	59.4	59.4	77.6	86.2	71.4	62.3	74.4
nli-distilbert-base-max-pooling	62.4	57.0	57.5	57.5	60.2	58.9	77.5	86.2	68.8	61.7	73.6
nli-distilroberta-base-v2	63.2	58.5	59.5	58.7	**61.5**	60.3	81.0	87.3	77.9	**71.5**	**79.4**
nli-mpnet-base-v2	64.2	58.8	59.7	59.1	60.6	60.5	81.0	87.2	78.1	69.3	78.9
nli-roberta-base	62.0	59.1	58.9	59.2	59.0	59.6	77.7	86.3	72.0	62.4	74.6
nli-roberta-large	63.9	59.5	**60.2**	**59.5**	61.3	**60.9**	80.7	87.2	77.7	70.9	79.1
nli-roberta-base-v2	62.8	59.7	58.9	59.3	60.8	60.3	77.9	86.5	72.0	63.1	74.9
paraphrase-distilroberta-base-v1	**65.0**	57.8	59.3	59.0	59.7	60.2	80.1	87.1	76.2	70.7	78.5
paraphrase-xlm-r-multilingual-v1	63.6	58.1	58.8	57.3	59.8	59.5	79.1	86.6	74.2	67.8	77.0
stsb-bert-base	64.0	59.1	57.7	58.1	60.6	59.9	78.1	86.5	72.4	63.4	75.1
stsb-bert-large	62.4	56.9	58.0	58.1	61.4	59.4	77.5	86.5	71.3	62.4	74.4
stsb-distilbert-base	62.8	58.0	58.0	57.1	59.3	59.1	78.5	86.5	73.1	64.6	75.7
stsb-distilroberta-base-v2	63.8	58.9	58.5	58.9	59.8	60.0	**81.1**	87.2	77.3	71.0	79.2
stsb-mpnet-base-v2	64.2	58.6	58.7	59.0	61.1	60.3	**81.1**	**87.5**	78.0	69.1	78.9
stsb-roberta-base	63.4	58.2	57.4	57.8	59.5	59.3	80.3	86.8	76.1	65.8	77.2
stsb-roberta-base-v2	63.4	58.7	59.7	58.9	60.6	60.3	81.0	87.3	77.5	70.3	79.0
stsb-roberta-large	62.7	58.4	57.6	58.0	59.7	59.3	80.1	86.6	74.2	65.4	76.6

TABLE IV
THE PEARSON CORRELATION BETWEEN THE PREDLABEL ACCURACY AND THE ESSAYS ACCURACY FOR ALL SENTENCE-BERT EMBEDDINGS. *P <.05. **P <.001, TWO-TAILED.

O	C	E	A	N	Ave.
0.086	0.488*	0.208	0.662**	0.533**	0.700**

Average results have shown that this model is successful in learning the personality trait-specific representations while retaining its knowledge from the pre-trained BERT. Even though the Bi-LSTM with max pooling outperforms the previous state-of-the-art when compared by performance metrics as well as richer personality trait-specific representation generation, the Sentence-BERT based model outperforms this one. We have experimented with different variants of Sentence-BERT. Among them, the most prominent results are found when RoBERTa-large or MPNet are used as the encoder in the Sentence-BERT architecture. In terms of accuracy of the *PredLabel*s and *SimScore*s, overall, RoBERTa-large performs the best. It achieves an accuracy for *PredLabel* of 97.65% which is almost double the previous state-of-the-art model's accuracy [41]. Apart from Agreeableness, its *PredLabel* accuracy is 100%, whereas for Agreeableness, it's 88.24%. MPNet achieves 100% *PredLabel* accuracy for Agreeableness. On average MPNet achieves 95.29% *PredLabel* accuracy. In terms of *SimScore*s, RoBERTa-large performs the best in all cases apart from Agreeableness. Still, its average value, 0.292, is more than three times that of [41]'s result.

For Agreeableness, the encoder with MPNet performs the best for *SimScore*, 0.305, and on average it achieves 0.233. Fig. 5 portrays in a 2D projection the representations generated by the RoBERTa-large version of Sentence-BERT, showing that the closeness of each statement to any particular trait is very clear. For each of the personality traits, the 0-, 1- and B-statements form distinguishable and well-separated clusters (with a couple of exceptions) as demonstrated. One issue of note, two metrics, *PredLabel* and *SimScore*, are used to measure the performance of the model. PCA has been used only to provide a visualization of the embeddings to show how close the representations of the similar trait samples are. We have also used other visualization techniques like t-SNE, UMAP, and LDA. As all the visualization results are very similar, we have reported only the PCA visualization.

To evaluate the generalizability of the model, we tested these models on the Essays and the Kaggle personality datasets. This time the Bi-LSTM with max-pooling performs worse than [41]'s work. The overall accuracy is almost 2% lower for the Essay dataset. But this is justifiable as this Siamese model was trained on very short sentences from the BFI statements, whereas the Essays dataset comes with long paragraphs. Additionally, LSTM based models face shortcomings while working with very long sequences. But the Sentence-BERT models, without any kind of additional operations, outperform the BERT-based averaging technique [41]. This time, RoBERTa-large achieves 60.9% accuracy which is an almost 1 percentage point boost compared to previous works.

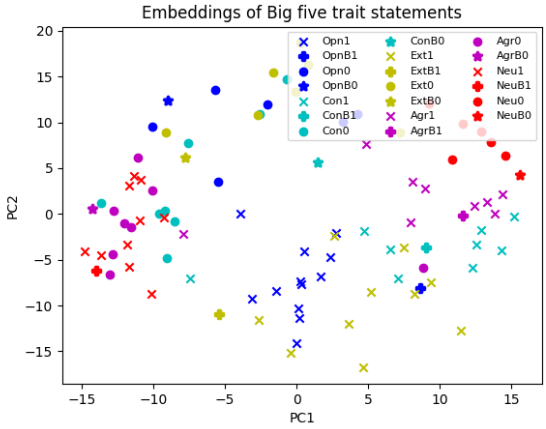

Fig. 5. Visualization of the personality statements after applying PCA on nli-roberta-large version of Sentence BERT. 1 and 0 mean "High" and "Low" rate of a specific trait, respectively, and "B" is for baseline sentences.

In the case of the Kaggle personality dataset, RoBERTa-large gains almost 2 percentage points more accuracy (79.1%). However, DistilRoBERTa performs the best for this dataset and achieves 79.4% accuracy. In both cases, MPNet shows prominent results with accuracies 60.3% and 78.9%, respectively. We also computed the Pearson correlation of the accuracy of PredLabel and Essays to see if the PredLabel accuracy gives any insight into how an encoder works for real world datasets. As demonstrated in Table V, although the experimented encoders are not specifically designed for long sequence datasets such as Essays, for most traits, especially the average of the traits, there is a significant positive correlation between these two accuracies. Hence, we can conclude that using PredLabel is a good approach for picking the best encoder for real-life datasets.

One notable significance of these models is that none of them have been enhanced with any kind of additional psychological features, unlike [41]. While training, the models are simply trained with sentence pairs. Thus it reduces the computational overhead as well. And as RoBERTa-large was initially trained over larger sequences and then fine-tuned again over natural language inference data, Sentence-BERT with RoBERTa-large earns the capability to produce sentence embeddings preserving richer semantics than the others. Furthermore, as the Sentence-BERT models are trained on a very large corpus of real life inference data compared to the siamese LSTM model, which is trained on the small BFI statement pairs dataset, they have achieved the ability to provide better representations of the statements.

VI. Conclusion

In this paper, we addressed the weaknesses of state-of-the-art personality detection models. With computationally less overhead, our model delivers sentence embeddings for psychological statements with rich semantics. Our results showed that our enriched representations distinguish the personality traits better than the CLS and average methods which are common in the field. Furthermore, we used the enriched representations in addition to Sentence-BERT models to classify traits based on their closeness to the baseline psychological statements so the result can be regarded as interpretable. Our experiments improved the Kaggle state-of-the-art accuracy by 2.3 percentage points and Essays by 0.3 percentage points. This work restricts the statements at the sentence level. In future, it can be extended to the paragraph level using hierarchical models like SMITH [58] so that better representations from the paragraphs can also be captured. Besides, BFI statements can be used within the prediction model to identify the closeness of each of the samples in the dataset with each of the BFI statements. We believe this method will help psychologists to get better insights into the prediction.

References

[1] P. Patalay and S. H. Gage, "Changes in millennial adolescent mental health and health-related behaviours over 10 years: a population cohort comparison study," *International Journal of Epidemiology*, vol. 48, no. 5, pp. 1650–1664, 2019.

[2] J. Orlowski, Jan 2021. [Online]. Available: https://www.thesocialdilemma.com/

[3] J. Achenbach, "Coronavirus is harming the mental health of tens of millions of people in U.S., new poll finds," Apr 2020. [Online]. Available: https://www.washingtonpost.com/health/coronavirus-is-harming-the-mental-health-of-tens-of-millions-of-people-in-us-new-poll-finds/2020/04/02/565e6744-74ee-11ea-85cb-8670579b863d_story.html

[4] A. Fowers and W. Wan, "A third of Americans now show signs of clinical anxiety or depression, Census Bureau finds amid coronavirus pandemic," May 2020. [Online]. Available: https://www.washingtonpost.com/health/2020/05/26/americans-with-depression-anxiety-pandemic/?arc404=true

[5] A. Samet, "2020 US social media usage: How the coronavirus is changing consumer behavior," Jun 2020. [Online]. Available: https://www.businessinsider.com/2020-us-social-media-usage-report

[6] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.

[7] "COVID-19 disrupting mental health services in most countries, WHO survey," Oct 2020. [Online]. Available: https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey

[8] D. J. Ozer and V. Benet-Martínez, "Personality and the prediction of consequential outcomes," *Annual Review of Psychology*, vol. 57, no. 1, pp. 401–421, 2006.

[9] B. W. Roberts, N. R. Kuncel, R. Shiner, A. Caspi, and L. R. Goldberg, "The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes," *Perspectives on Psychological Science*, vol. 2, no. 4, pp. 313–45, 2007.

[10] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative Big Five Trait taxonomy," *Handbook of Personality: Theory and Research*, pp. 114–158, 2008.

[11] D. P. Goldberg and P. Huxley, *Common mental disorders: a bio-social model.* Tavistock/Routledge, 1992.

[12] L. Campbell-Sills, S. L. Cohan, and M. B. Stein, "Relationship of resilience to personality, coping, and psychiatric symptoms in young adults," *Behaviour Research and Therapy*, vol. 44, no. 4, pp. 585–599, 2006.

[13] H. Yin, Y. Wang, Q. Li, W. Xu, Y. Yu, and T. Zhang, "A network-enhanced prediction method for automobile purchase classification using deep learning." in *PACIS*, 2018, p. 111.

[14] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell, "Psychological targeting as an effective approach to digital mass persuasion," *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, pp. 12 714–12 719, 2017.

[15] S. Dhelim, N. Aung, M. Amine Bouras, H. Ning, and E. Cambria, "A survey on personality-aware recommendation systems," *Artificial Intelligence Review*, 2021.

[16] H. Schneider, K. Schauer, C. Stachl, and A. Butz, "Your data, your vis: Personalizing personal data visualizations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10515, 2017, pp. 374–392.

[17] S. T. Völkel, R. Schödel, D. Buschek, C. Stachl, Q. Au, B. Bischl, M. B"uhner, and H. Hussmann, "Opportunities and Challenges of Utilizing Personality Traits for Personalization in HCI: Towards a shared perspective from HCI and Psychology," in *Personalized Human-Computer Interaction*. Oldenbourg, Germany: De Gruyter, 2019.

[18] C. C. Liem, M. Langer, A. Demetriou, A. M. Hiemstra, A. S. Wicaksana, M. P. Born, and C. J. König, "Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pp. 197–253.

[19] J. Maria Balmaceda, S. Schiaffino, and D. Godoy, "How do personality traits affect communication among users in online social networks?" *Online Information Review*, vol. 38, no. 1, pp. 136–153, 2014.

[20] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[21] M. Kaya and H. S. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, p. 1066, 2019.

[22] O. P. John, L. Naumann, and C. & Soto, "Paradigm shift to the integrative big five taxonomy: History, measurement, and conceptual issues," in *Handbook of Personality: Theory and Research*. The Guilford Press, 2008, pp. 114–158.

[23] O. P. John, E. M. Donahue, and R. L. Kentle, "Big Five Inventory," *Journal of Personality and Social Psychology*, 1991.

[24] V. Benet-Martínez and O. P. John, "Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English." *Journal of Personality and Social Psychology*, vol. 75, no. 3, p. 729, 1998.

[25] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of Personality and Social Psychology*, vol. 77, no. 6, p. 1296, 1999.

[26] M. Jolly, "(MBTI) Myers-Briggs personality type dataset," Sep 2017. [Online]. Available: https://www.kaggle.com/datasnaek/mbti-type

[27] D. J. Pittenger, "Cautionary comments regarding the Myers-Briggs type indicator." *Consulting Psychology Journal: Practice and Research*, vol. 57, no. 3, p. 210, 2005.

[28] R. J. Gregory, "The History of Psychological Testing," in *Psychological Testing: History, principles, and applications.*, 7th ed. Pearson, 2013, ch. 2, pp. 32–58.

[29] I. Briggs Myers, *Introduction to Type: A Guide to Understanding Your Results on the Myers-Briggs Type Indicator (revised by L K Kirby & K D Myers)*. CA: Consulting Psychologists Press, 1993.

[30] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proceedings of the National Academy of Sciences*, vol. 112, no. 4, pp. 1036–1040, 2015.

[31] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[32] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[33] Y. Li, A. Kazameini, Y. Mehta, and E. Cambria, "Multitask learning for emotion and personality detection," *arXiv preprint arXiv:2101.02346*.

[34] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, pp. 2313–2339, 2019.

[35] C. Stachl, R. L. Boyd, K. T. Horstmann, P. Khambatta, S. Matz, and G. M. Harari, "Computational personality assessment-an overview and perspective," *PsyArXiv*, 2021.

[36] R. Lewis, *When Cultures Collide*. Boston, MA: Nicholas Brealey Publishing, 2010.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[38] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[41] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 1184–1189.

[42] P. Vijayaraghavan, E. Chu, and D. Roy, "DAPPER: Learning domain-adapted persona representation using pretrained BERT and external memory," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 643–652.

[43] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality trait detection using bagged SVM over BERT word embedding ensembles," in *Proceedings of the The Fourth Widening Natural Language Processing Workshop*. Association for Computational Linguistics, 2020.

[44] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.

[45] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[46] D. G. Johnson, *Computer Ethics*. Englewood Cliffs (NJ): Wiley Online Library, 1985.

[47] H. Grassegger and M. Krogerus, "The data that turned the world upside down," *Vice Motherboard*, vol. 28, 2017.

[48] P. T. Costa Jr and R. R. McCrae, "Domains and facets: Hierarchical personality assessment using the revised neo personality inventory," *Journal of Personality Assessment*, vol. 64, no. 1, pp. 21–50, 1995.

[49] L. R. Goldberg, "International personality item pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences," *Retrieved August*, vol. 14, p. 2002, 1999.

[50] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[51] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[52] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[53] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[54] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[55] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," *arXiv preprint arXiv:1704.05426*, 2017.

[56] N. Reimers, "Pretrained models," 2020. [Online]. Available: https://www.sbert.net/docs/pretrained_models.html

[57] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," *arXiv preprint arXiv:2004.09297*, 2020.

[58] L. Yang, M. Zhang, C. Li, M. Bendersky, and M. Najork, "Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1725–1734.