Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17015-z

DLP-personality detection: a text-based personality detection framework with psycholinguistic features and pre-trained features

Hao Lin[1]

Received: 9 September 2022 / Revised: 15 August 2023 / Accepted: 11 September 2023

Abstract
Text-based personality detection aims to identify the personality traits implied in subject-supplied textual data. However, over-reliance on pre-trained language models and neglect of psycholinguistic features has become a bottleneck in personality detection. In this work, we conduct extensive feature-level ablation experiments using multiple psycholinguistic features to verify the importance of psycholinguistic features for personality detection. Furthermore, we propose a novel personality detection framework, DLP-Personality Detection, which fuses multiple psycholinguistic features and pre-trained language features. With the DLP-Personality Detection, we achieve state-of-the-art performance for the Big Five personality traits (Big 5) and Myers-Briggs Type Indicator (MBTI) personality traits on the Essays and Kaggle MBTI datasets. Finally, we provide some suggestions for psycholinguistic features and discuss future work for personality detection.

Keywords Personality detection · Psycholinguistic features · Pre-trained language features · Feature fusion

1 Introduction

Personality is a psychological construction that has been associated with a wide range of crucial life outcomes and choices of people [1]. Any technology involving understanding, prediction, and synthesis of human behavior, such as human-computer interaction [2], recommended system [3], and mental illness diagnosis [4], is likely to benefit from personality detection which is a fundamental task in psychology.

Traditional manual measurement approaches of personality, such as the Self-report Inventory, are widely used by psychology scholars, but abandoned by computer science scholars due to their low efficiency. The rejections spawned machine learning-based methods of automatic personality detection, which dominate today. These machine learning-based detection

✉ Hao Lin
suzukaze_aoba@126.com

[1] Tianjin University of Technology, School of Computer Science and Engineering, Tianjin, China

Published online: 29 September 2023

methods are trained with texts [5–7], images [8], audio [9], videos, social application data [10], and even electroencephalograms [11]. In particular, the text-based detection method is the cornerstone of other sophisticated detection methods and the most practical because the text data about subjects are easily accessible.

Transfer Learning (TL) has proved to be overwhelmingly useful in the data mining field since most fields typically do not have millions of labeled real-world data to train complex models, as is the field of personality detection. However, we find that the existing text-based personality detection methods rely too much on pre-trained language models of TL, such as BERT [5], XLNet [12], and Robert [6] to omit key psycholinguistic features. Individual differences in linguistic utilization have been considered as reflections of psychological phenomena since the early days of Freud [5]. The choice of words is driven not only by meaning, but also by psychological phenomena such as emotions, relational attitudes, power status, and personality traits [1, 13]. So, these psycholinguistic features are also significant for personality detection and higher model interpretability.

To address the above-mentioned problems, we conduct extensive feature-level ablation experiments using multiple psycholinguistic features and publicly available datasets including Essays and Kaggle MBTI. The psycholinguistic features adopted by us contain almost all psycholinguistic features that have been used for personality detection. Furthermore, we propose a novel text-based personality detection framework with multiple psycholinguistic features and pre-trained language features, called DLP-Personality Detection. Benefit from the results of the ablation experiments, we verified the effectiveness of this framework. The main contributions of this paper can be summarized as follows:

- We propose the DLP-Personality Detection, which consists of a preprocessing module, a feature extraction module, and a classifier module. The framework adopts multiple psycholinguistic features and pre-trained language features as inputs and a Bi-LSTM with an attention mechanism as a classifier.
- We conduct extensive feature-level ablation experiments for Big 5 and MBTI personality traits-related datasets to verify that psycholinguistic features should not be ignored for personality detection. As far as we know, it is the first time to construct a text-based personality detection framework with such high-dimensional psycholinguistic features.
- We verify that the DLP-Personality Detection achieves state-of-the-art performance for personality detection by using the publicly available Essays and Kaggle MBTI dataset.

2 Relate work and preliminaries

2.1 Personality taxonomies

Personality theory is divided into six schools of psychoanalysis, traits, biology, humanism, behaviorism, and cognition. With time going by, numerous taxonomies for the description of human personality have been proposed. At present, the most representative and frequently used taxonomies are the Big 5 and MBTI in the personality trait school, which are shown in Figs. 1 and 2 [14].

Big 5, a consensus among researchers on personality description, is constructed by the lexical method and describes the individual's personality from five personality traits. The MBTI emphasizes naturally occurring differences more and indicates people's differing psychological priorities in perceiving the real world and making decisions. As shown in Figs. 1 and 2,

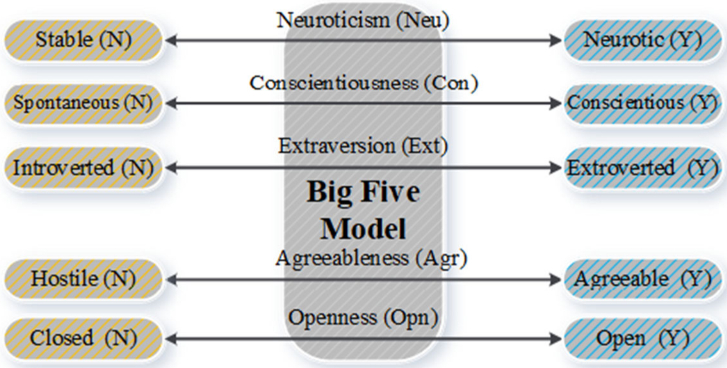

Fig. 1 Big Five personality taxonomies

the detection of each personality trait can be regarded as a binary classification problem [15]. It is the mainstream personality detection mode.

In addition, other personality taxonomies such as the Minnesota Multiple Personality Inventory (MMPI), sixteen personality factor questionnaire (16PF), and Eysenck Personality Questionnaire (EPQ) are widely used in psychology. However, due to the lack of relevant public datasets, there is no relevant personality detection research.

2.2 Text-based personality detection using psycholinguistic features

One of the early efforts in personality detection was proposed by [16]. The words in corpora were grouped into four psychologically meaningful categories: function, cohesion, assessment, and appraisal. The detection task was performed with a SVM, whose input was the relative frequencies of the words appearing in each category. Mairesse et al. used the same corpora and SVM, but extra adopted LIWC and Medical Research Council (MRC) psycholinguistic features to achieve an average accuracy of 57% [15]. Nguyen et al. labeled 10000

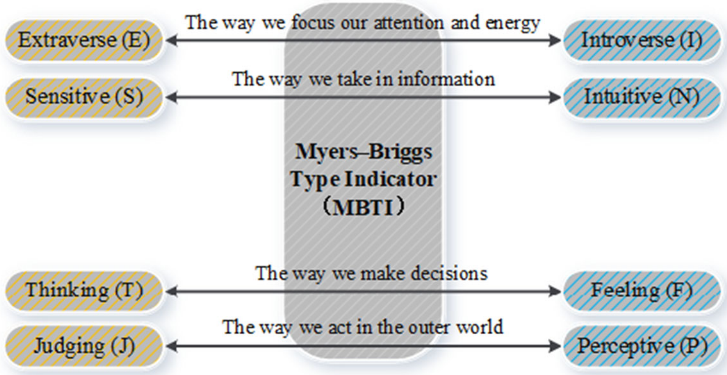
The way we focus our attention and energy
Extraverse (E)
Introverse (I)
The way we take in information
Sensitive (S)
Intuitive (N)
Myers-Briggs Type Indicator (MBTI)
The way we make decisions
Thinking (T)
Feeling (F)
The way we act in the outer world
Judging (J)
Perceptive (P)

Fig. 2 Myers-Briggs Type Indicator

users of Livejournal and adopted LIWC features and logistic regression to detect the Big 5 personality traits of users [17]. Poria et al. proposed a more sophisticated detection method whose inputs are LIWC, MRC, and SenticNet features [18]. Moreover, these features were used to build a SMO classifier.

2.3 Text-based personality detection using pre-trained language models

Mehta et al. reported their results on Essays and Kaggle MBTI dataset with two pre-trained language models including BERT-base and BERT-large [7]. They believe that their detection model consisting of BERT and MLP dominated the detection of the Big 5 and MBTI personality traits and the features extracted by pre-trained language models consistently beat conventional psycholinguistic features. Wang et al. proposed a deep learning-based framework for personality detection from text data with Capsule Networks and XLNets [12]. Jiang et al. presented a novel approach to automatic personality detection using the pre-trained language model RoBERT and attentive neural networks [6]. Their model improves the state-of-art results on the Essays dataset by 2.49%. Likewise, Ricardo et al. proposed a personality detection approach with RoBERT for the MBTI personality traits [19]. Kamal et al. [20] used three pre-trained models including Elmo, ULMFiT, and BERT to extract features and achieved SOTA results on the myPersonality dataset. Likewise, Lopez et al. [21] adopted Word2Vec, GloVe, and BERT to detect personality traits.

In recent years, the combination of psycholinguistic features and pre-trained language features has been increasingly used for personality detection. Kazameini et al. concatenated features extracted by BERT with the Mairesse features, which are made up of LIWC, MRC, prosodic and utterance-type features [5]. They fed these features to multiple SVMs to detect personality traits in parallel like a bagging classifier. Similarly, Ren et al. leverage BERT and SenticNet 5 features to detect personality from text data [22].

To make a long story short, the above works only use pre-trained language models or a few psycholinguistic features additionally. Individual differences in linguistic utilization have been considered as reflections of psychological phenomena. The psycholinguistic features have the same significance for personality detection as pre-trained language models and more interpretability. Apart from Mairesse features and SenticNet 5 features, other psycholinguistic features such as NRC Emotion Lexicon features, NRC VAD Lexicon features, Hourglass of Emotions features, and text readability features have been proved to relate to personality traits by correlation analysis or factor analysis. These features should be given more attention for personality detection.

3 Methodology

In this section, we introduce our DLP-Personality Detection framework in detail. As shown in Fig. 3, the framework is divided into a preprocessing module, a feature extraction module, and a classifier module.

3.1 Preprocessing module

In the preprocessing module, we augment samples that belong to label-low classes to mitigate the impact of sample imbalance on the generalization of training models. We use Easy

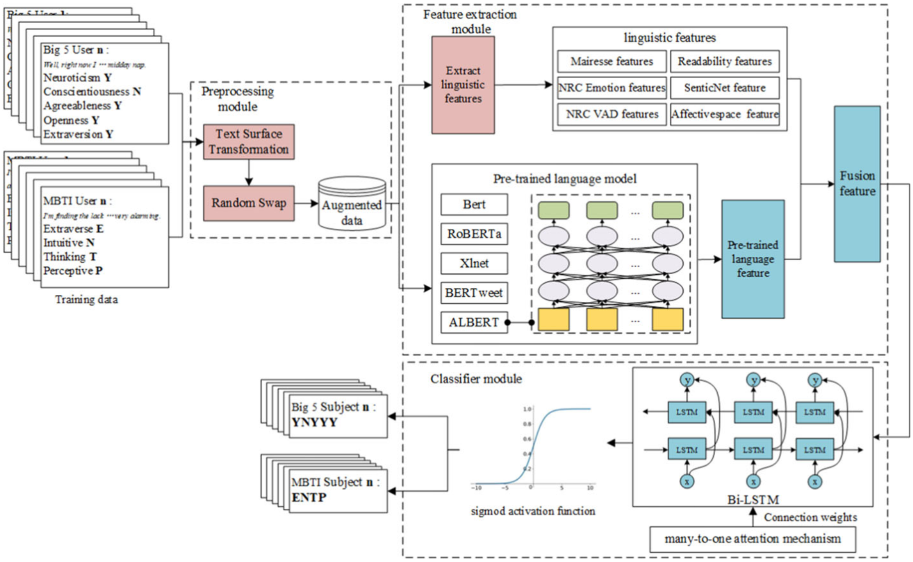

Fig. 3 DLP-Personality Detection proposed by us. “DLP” presents three key parts of our framework: data augmentation, psycholinguistic features, and pre-trained language features

Data Augmentation (EDA) to enhance samples of label-low classes [23]. In order to extract psycholinguistic features accurately, we perform Text Surface Transformation (TST) on each sample before using the above operations [24]. Personality detection can be regarded as multiple binary-classification problems, owing to this, we binary-coded the personality traits into 0, 1.

Moreove, It is not a good idea to augment the entire dataset, since the sample imbalance will be amplified. Furthermore, should not be too large to avoid losing more original semantics.

3.2 Feature extraction module

Benefiting from previous research, apart from the common Mairesse and SenticNet 5 features, we additionally adopt multiple psycholinguistic features for our DLP-Personality Detection. These psycholinguistic features consist of the following:

- The Mairesse has a set of psycholinguistic features consisting of LIWC, MRC, prosodic and utterance-type features [15]. We abandoned prosodic features and finally adopted a total of 79 features. These are the widely used features in traditional machine learning-based personality trait mining.
- The NRC Emotion Lexicon has a lexicon of over 14,000 English words which are annotated with values of emotions such as anger, anticipation, disgust, etc [25]. The final value of this sub-feature is the means of all values of emotionally charged words present in the text data.
- The NRC VAD Lexicon has a lexicon of over 20,000 English words which is annotated with their valence, arousal, and dominance scores [26]. As above, the VAD Lexicon value is the means of all constituent words in the text data.

- The Affectivespace is a vector space of affective common sense available for English and has 100,000 concepts [27].
- The Readability[1] has a number of calculated readability measures which are based on simple surface characteristics of the text data. These measures are basically linear regressions based on the number of words, syllables, and sentences.
- The SenticNet 5 is a tool used for extracting common sense knowledge along with associated sentiment polarity and affective labels from the text data, including pleasantness value, attention value, sensitivity value, aptitude value, and polarity value [28].

The above features contain almost all psycholinguistic features that have been used for personality trait mining. The relationship between these features and personality traits has been discussed in a multitude of literature.

The good representation obtained by pre-trained language models express general-purpose priors that are not task-specific but would be useful for a learning machine to solve data mining tasks [29]. When it comes to language, a good representation should capture the implicit linguistic rules and common sense knowledge hiding in text data, such as lexical meanings, syn- tactic structures, semantic roles, and even pragmatics. We adopt multiple pre-trained language models for our DLP-Personality Detection including BERT, RoBERT, Xlnet, ALBERT, and BERTweet to carry out language model embeddings. Take the latest ALBERT as an example, which is a pre-trained language model proposed by Google Research. In ALBERT, $[E_1, E_2, \ldots, E_n]$ represents the original text. Trm means the Transformer. Finally, $[T_1, T_2, \ldots, T_n]$ is output as the feature representation vector of $[E_1, E_2, \ldots, E_n]$, which contains entire text information of text sequence.

We employ an early fusion method that combines psycholinguistic features with pre-trained language features as

$$T = concat([Mairesse, NRC, \ldots, SenticNet], [T_1, T_2, \ldots, T_n]) \tag{1}$$

3.3 Classifier module

The classifier module is responsible for using the fused features to classify personality traits. We build a single-layer Bi-LSTM with 128 units for our DLP-Personality Detection. Bi-LSTMs, improved from LSTM, put two independent LSTMs together. This structure allows Bi-LSTMs to have both backward and forward information about the sequence. The forward information $\vec{h}_t$ can be calculated as

$$f_t = sigmoid(W_f[h_(t-1), x_t] + b_f), \tag{2}$$

$$i_t = sigmoid(W_i[h_(t-1), x_t] + b_i), \tag{3}$$

$$C'_t = tanh(W_C[h_(t-1), x_t] + b_C), \tag{4}$$

$$C_t = f_t C_(t-1) + i_t C'_t, \tag{5}$$

$$o_t = sigmoid(W_o[h_(t-1), x_t] + b_o), \tag{6}$$

$$\vec{h}_t = o_t tanh(C_t). \tag{7}$$

At each time t, x_t is the current input. f_t, i_t, C_t, o_t represent the forget gate, input gate, cell state, and output gate, respectively. W means the weight of gates. b means the bias. Conversely, we get the backward information $\vec{h}_t$ and calculate the whole output h_t as

$$h_t = sigmoid(W_h[\vec{h}_t, \overleftarrow{h_t}] + b_h) \tag{8}$$

[1] pypi.org/project/readability/

Additionally, we also experiment with SVM, LR, RF, MLP, LSTM, and multi-layer Bi-LSTM while fine-tuning, however, it results in no evident performance boost.

Binary Cross Entropy is used as the loss function of the Bi-LSTM. For the personality class label y_i and the total number of samples N, the Bi-LSTM is trained using

$$CE = \frac{1}{N}\sum_{i=1}^{N} y_i \log(p(y_i)) + (1 - y_i)\log(1 - p(y_i)) \tag{9}$$

In addition, we introduce a many-to-one attention mechanism into Bi-LSTM to weight fuse the output vectors to make the contribution distribution of output vectors more reasonable. The attention function of the many-to-one attention mechanism is

$$Attention(h_t, h_s) = \frac{exp(Socre(h_t, h_s))}{\sum_{s'=1}^{s} exp(Socre(h_t, h'_s))} \tag{10}$$

We adopt the Dot function as Socre function. Additionally, we experiment with more complex attention mechanisms, such as Scaled Dot-product Attention, Bahdanau Attention, and Multi-Head Attention, yet they result in a performance drop.

The output results of the Bi-LSTM are input to a Dense layer with a Sigmoid activation function to normalize the personality trait results, i.e.,

$$Confidence = \frac{1}{1 + exp^{-h_t}} \tag{11}$$

If the confidence coefficient of a trait is greater than the probability threshold of 0.5, it is considered that the trait belongs to a category, resulting in a trait label of Y. Conversely, if the confidence coefficient of a trait is less than the probability threshold of 0.5, it is considered that the trait does not belong to a category, resulting in a trait label of N.a

4 Experiments and discussion

4.1 Experiment datasets

We adopt the publicly available Essays and Kaggle MBTI datasets in our experiments. Essays, a scientific gold standard from psychology for personality detection, consists of 2468 essays written by students and annotated with the binary labels of the Big 5 personality traits which were identified by a standardized Self-report Inventory. Kaggle MBTI contains tweets posted by 8675 users and was labeled by a Self-report Inventory. For the Essays dataset, we augment a total of 430 new samples belonging to 23 label-low classes. is set to 0.05. And we do not augment the new samples for the Kaggle MBTI dataset.

4.2 Evaluation metrics

We utilize accuracy, recall, and F1 as the primary performance metrics. Among them, accuracy was the most widely adopted in numerous related works.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN). \tag{12}$$

$$Recall = TP/(TP + FP). \tag{13}$$

$$Precision = TP/(TP + FN). \tag{14}$$

$$F1 = 2 \times PrecisionRecall/(Precision + Recall). \tag{15}$$

Where TP indicates the number of the positive samples which are correctly classified. TN is the number of negative samples which are correctly classified. FP represents the number of negative samples which are wrongly classified. FN is the number of positive samples which are wrongly classified.

4.3 Feature-level ablation experiments

In order to explore the effects of different psycholinguistic features, we employ several sets of comparison experiments by fixing model structures and hyper parameters, and only changing the testing psycholinguistic features. These fixed model structures and hyper parameters are as follows: the most common five-layer MLP with 256x128x64x32x2 units, RELU activation function, Adam optimizer, 0.0001 learning rate, 64 batch size, 50 epochs, 10 fold cross-validation. Each set of testing psycholinguistic features will be combined with the pre-trained language features extracted by Albert-large through (1). The experiment dataset is the unaugmented Essays and Kaggle MBTI dataset. These results of features-level ablation experiments serve as the basis for the decision and recommendations of the final features of the DLP-Personality Detection.

In order to facilitate table drawing, we omit Mairesse feature whose importance is indisputable and abbreviate the psycholinguistic features as follows: NRC Emotion Lexicon = N; NRC VAD Lexicon = V; Affectivespace = A; Readability = R; SenticNet 5 = S. The experiment results are shown in Tables 1 and 2.

From the experimental results in Tables 1 and 2, we give several conclusions and suggestions as follows: 1) The personality detection model only using pre-trained language features achieves the highest accuracy. This is consistent with the state-of-the-art accuracy reported in much literature. However, the model using only pre-trained language features achieves low recall, and low F1 value. Much literature did not report recall and F1 values of their methods. So, we suggest using multiple features to detect personality traits. 2) The more features are used, the higher the models' performance may not be. We suggest targeted selecting psycholinguistic features for different personality taxonomies. 3) The Readability is a set of important features for recall and F1 values of the detection models. We suggest using the Readability feature to detect personality traits.

These low recall and F1 values are mainly caused by the severe sample imbalance in the unaugmented dataset. In addition, due to the Social Desirability Effect [30, 31], the subjects tend to be biased toward the good side when they are conducting Self-report Inventories. It may cause subjects to give a misleading answer, resulting in severe sample imbalance and textual features do not resonate well with personality traits from questionnaires [32].

The experiment results on the Kaggle MBTI dataset are similar to that on the Essays dataset. We also experiment with deeper classifiers, however, it results in similar results. We believe that the misunderstanding of over-reliance on pre-trained language models will be avoided with this experiment. Finally, we comprehensively consider the accuracy and F1 in Tables 1 and 2 to select psycholinguistic features. For the Big 5, we adopt Mairesse, NRC Emotion Lexicon, NRC VAD Lexicon, Affectivespace, Readability, and SenticNet5 features. For the MBTI, Mairesse, NRC Emotion Lexicon, NRC VAD Lexicon, and Readability features are adopted for follow-up experiments.

Table 1 Feature ablation experiment about Big 5 personality traits on essays dataset

Features	Accuracy						Recall					F1				
	Avg	EXT	NEU	AGR	CON	OPN	EXT	NEU	AGR	CON	OPN	EXT	NEU	AGR	CON	OPN
N	54.91	54.07	54.76	55.33	55.25	55.13	18.74	21.81	19.83	29.39	28.04	28.03	30.25	29.32	36.21	35.57
V	54.71	54.40	54.60	55.53	54.28	54.76	15.89	23.50	19.15	23.64	24.57	25.89	31.64	28.71	32.10	32.61
A	54.70	54.16	54.32	55.09	54.64	55.29	13.75	22.48	16.13	23.73	25.65	24.00	30.59	26.13	31.96	33.63
S	54.93	54.03	54.92	55.25	55.05	55.37	16.07	22.70	19.82	25.16	27.49	25.81	31.11	29.18	33.30	34.89
R	56.51	55.94	55.62	56.51	56.14	58.37	50.22	47.45	48.69	51.96	56.48	53.14	51.03	51.77	54.36	57.71
N+V	54.81	53.91	54.60	55.37	55.17	55.01	17.75	24.64	20.74	28.12	26.66	27.38	32.67	29.90	35.44	34.44
N+A	54.94	54.64	54.72	54.76	55.17	55.41	17.52	24.17	21.75	27.61	28.94	27.13	31.95	30.65	35.08	36.12
N+S	54.58	53.63	54.56	55.33	54.88	54.52	18.06	24.26	20.13	30.06	28.8	27.77	32.23	29.08	37.23	35.93
N+R	56.55	55.70	55.86	56.87	55.70	58.61	51.90	48.79	52.14	53.17	**57.74**	54.26	51.69	54.50	55.19	58.50
V+A	54.78	54.16	54.76	55.41	54.52	55.05	11.52	22.16	18.05	24.16	26.09	21.73	30.53	27.64	32.29	33.95
V+S	55.09	54.52	55.29	55.45	55.01	55.17	14.11	22.28	17.79	24.91	27.22	24.38	30.80	27.40	32.84	34.96
V+R	56.40	55.33	55.65	56.55	56.06	58.41	49.83	48.85	49.12	51.53	55.34	52.64	51.86	52.19	53.82	56.85
A+S	54.95	54.60	54.76	55.17	55.37	54.85	17.01	23.31	16.76	24.34	28.25	26.57	31.22	26.6	32.48	35.50
A+R	56.35	55.90	55.29	56.30	56.18	58.09	49.24	46.94	48.58	50.53	55.27	52.33	50.41	51.69	53.09	56.65
S+R	56.21	55.29	55.17	56.10	56.22	58.25	51.86	47.45	49.35	51.67	55.39	54.16	50.85	52.15	53.88	56.85
N+V+A	54.82	53.99	54.60	55.05	55.17	55.29	15.67	22.57	20.84	28.05	27.14	25.51	30.74	29.58	35.46	34.55
N+V+S	55.08	54.28	54.89	56.02	55.05	55.17	16.05	23.22	19.53	27.76	28.29	25.88	31.05	28.64	35.04	35.59
N+V+R	56.48	56.02	55.33	56.22	55.98	58.86	51.16	49.18	49.57	52.21	56.67	53.81	52.16	52.65	54.37	57.74
N+A+S	54.97	54.28	54.56	55.17	55.45	55.41	17.24	23.86	19.96	28.15	30.61	26.63	31.98	29.19	35.32	37.31
N+A+R	56.54	55.41	56.18	56.26	56.27	58.57	49.58	46.82	50.68	52.30	55.21	52.54	50.37	53.18	54.40	56.64
N+S+R	56.63	56.31	55.37	56.30	56.34	58.82	52.20	48.23	49.35	52.91	57.13	54.50	51.34	52.29	54.99	58.19
V+A+S	55.01	54.20	55.45	55.21	55.09	55.13	15.35	23.21	16.07	23.44	27.83	25.19	31.68	25.84	31.83	35.24
V+A+R	56.12	56.02	55.17	55.86	55.17	58.37	50.45	47.13	49.11	51.74	55.83	53.13	50.42	52.14	54.11	57.19

Table 1 continued

Features	Accuracy						Recall					F1				
	Avg	EXT	NEU	AGR	CON	OPN	EXT	NEU	AGR	CON	OPN	EXT	NEU	AGR	CON	OPN
V+S+R	56.48	55.74	55.41	56.39	56.67	58.21	50.37	46.56	49.19	50.98	55.64	53.19	50.13	52.30	53.48	57.11
A+S+R	56.64	55.82	55.74	56.38	56.55	58.69	51.15	46.60	48.44	50.91	55.69	53.69	50.30	51.72	53.42	57.08
N+V+A+S	55.23	54.44	54.89	55.25	55.33	56.26	15.68	24.28	19.63	29.00	29.18	25.89	32.05	28.95	35.99	36.33
N+V+A+R	56.63	55.86	**56.42**	56.22	56.06	58.57	50.17	46.93	49.34	51.61	55.39	52.89	50.43	52.22	53.83	56.77
N+V+S+R	56.61	55.94	55.86	56.26	56.43	58.57	50.97	48.73	49.28	**53.30**	56.98	53.57	51.90	52.30	55.21	58.00
N+A+S+R	56.90	55.74	56.18	56.51	57.11	58.98	50.20	46.66	50.27	52.10	55.55	52.86	50.24	52.94	54.24	56.97
V+A+S+R	56.46	55.37	55.41	56.34	56.59	58.57	48.13	46.05	49.37	51.20	56.03	51.31	49.89	52.37	53.67	57.29
N+V+A+S+R	56.43	55.49	55.33	56.02	57.03	58.25	**55.41**	**52.09**	**52.69**	51.82	53.53	**54.63**	**55.09**	**55.30**	**56.57**	**58.23**
Not adopted	**57.99**	**57.15**	55.53	**58.37**	**59.22**	**59.67**	42.47	38.71	38.55	41.36	42.43	47.42	44.70	44.40	46.64	47.12

Table 2 Feature ablation experiment about MBTI personality traits on Kaggle MBTI dataset

Features	Accuracy					Recall					F1				
	Avg	I/E	N/S	T/F	P/J	Avg	I/E	N/S	T/F	P/J	Avg	I/E	N/S	T/F	P/J
N	73.12	77.24	86.25	67.62	61.38	60.66	66.95	80.41	53.39	41.9	66.14	68.93	81.84	61.66	52.15
V	72.95	77.29	86.26	67.07	61.20	60.05	66.95	80.42	51.95	40.86	65.63	68.73	81.89	60.54	51.35
A	72.89	77.20	86.29	66.93	61.14	59.99	66.95	80.42	51.88	40.7	65.49	68.6	81.7	60.5	51.16
S	72.88	77.27	86.25	66.74	61.24	60.14	66.95	80.42	52.57	40.61	65.74	68.92	81.85	61.08	51.13
R	73.09	77.24	86.27	67.67	61.18	70.46	**72.78**	84.38	66.07	59.19	72.59	76.78	85.74	67.1	60.74
N+V	73.15	77.27	86.28	**67.77**	61.27	60.51	66.95	80.41	53.03	41.65	66.02	68.87	81.86	61.4	51.97
N+A	73.05	77.20	86.27	67.52	61.20	60.47	66.95	80.42	53	41.5	65.97	68.92	81.81	61.36	51.79
N+S	72.99	77.23	86.27	67.25	61.22	60.54	66.95	80.42	53.26	41.53	66.06	68.92	81.87	61.57	51.87
N+R	73.05	77.20	**86.29**	67.52	61.20	60.47	66.95	80.42	53	41.5	65.97	68.92	81.81	61.36	51.79
V+A	72.91	77.33	86.26	66.84	61.20	60.07	66.95	80.42	52.09	40.81	65.55	68.54	81.75	60.63	51.29
V+S	72.91	77.22	86.26	66.93	61.22	60.14	66.95	80.42	52.21	40.96	65.70	68.81	81.91	60.74	51.35
V+R	73.18	77.35	86.26	67.77	61.33	70.35	72.19	**84.41**	66.31	58.49	72.49	76.69	85.77	67.37	60.13
A+S	72.88	77.24	86.27	66.78	61.22	59.98	66.95	80.42	52.24	40.3	65.45	68.51	81.77	60.72	50.79
A+R	72.94	77.30	86.26	67.69	60.51	69.07	70.55	83.01	64.51	58.21	72.4	76.91	85.4	67.5	59.79
S+R	73.09	77.26	86.25	67.49	61.37	69.43	71.33	82.3	65.35	58.73	72.24	76.78	85.67	67.33	59.18
N+V+A	73.11	77.23	86.26	67.68	61.27	60.42	66.95	80.41	53.28	41.03	65.89	68.79	81.73	61.59	51.46
N+V+S	73.11	77.29	86.21	67.62	61.33	60.59	66.95	80.42	53.49	41.53	66.04	68.79	81.82	61.67	51.87
N+V+R	**73.20**	**77.36**	86.26	67.71	**61.46**	**70.84**	72.42	84.16	**66.58**	**59.61**	**73.02**	**77.08**	**85.78**	**67.52**	**61.7**
N+A+S	72.77	76.96	86.21	67.01	60.91	56.62	67.15	80.42	49.65	29.25	61.47	67.06	80.41	57.58	40.83
N+A+R	72.82	76.96	86.22	66.80	61.31	62.90	67.14	80.39	58.47	45.61	66.75	67.05	80.4	64.49	55.04

Table 2 continued

Features	Accuracy					Recall					F1				
	Avg	I/E	N/S	T/F	P/J	Avg	I/E	N/S	T/F	P/J	Avg	I/E	N/S	T/F	P/J
N+S+R	72.77	76.96	86.21	67.01	60.91	56.62	67.15	80.42	49.65	29.25	61.47	67.06	80.41	57.58	40.83
V+A+S	72.69	76.96	86.21	66.62	60.98	55.77	67.13	80.42	47.91	27.6	60.76	67.05	80.42	56.26	39.33
V+A+R	72.80	76.96	86.22	66.85	61.15	62.68	67.16	80.43	58.12	45.02	66.52	67.07	80.43	64.07	54.51
V+S+R	73.01	76.96	86.21	67.70	61.18	63.04	67.15	80.41	58.44	46.23	66.88	67.07	80.41	64.4	55.64
A+S+R	72.84	77.01	86.21	66.80	61.33	62.67	67.17	80.41	58.92	44.19	66.56	67.07	80.4	64.73	54.04
N+V+A+S	72.84	76.96	86.21	67.20	60.98	56.82	67.14	80.4	49.61	30.14	61.67	67.05	80.4	57.56	41.69
N+V+A+R	72.91	76.96	86.22	67.15	61.30	63.14	67.16	80.42	59.21	45.75	66.92	67.07	80.41	64.96	55.26
N+V+S+R	72.81	76.96	86.22	66.85	61.21	63.19	67.15	80.41	59.77	45.44	66.99	67.06	80.43	65.41	55.07
N+A+S+R	72.91	76.96	86.24	67.14	61.31	63.16	67.16	80.44	58.82	46.23	66.95	67.06	80.43	64.71	55.59
V+A+S+R	73.00	76.96	86.21	67.55	61.28	63.02	67.16	80.41	58.38	46.08	66.79	67.06	80.41	64.31	55.39
N+V+A+S+R	72.89	76.97	86.21	67.03	61.33	63.15	67.15	80.42	58.61	46.4	66.96	67.06	80.41	64.63	55.73
Not adopted	72.59	76.96	86.22	66.33	60.85	55.84	67.14	80.41	47.58	28.24	60.87	67.05	80.41	56.01	40.03

Table 3 Detail of pre-trained language models

Model	Layers	Hidden	Token length
BERT-base	12	768	512
BERT-large	24	1024	512
RoBERTa	12	768	512
Xlnet-base	12	768	512
Xlnet-large	24	1024	512
ALBERT-base	12	768	512
ALBERT-large	24	1024	512
BERTweet	12	768	128

4.4 Performance comparison with multiple pre-trained language models

In order to explore the effects of different pre-trained language models, we employ several performance comparison experiments by fixing model structures, hyper parameters, and psycholinguistic features. The details of the above pre-trained models are shown in Table 3. For each model, we optimized the hyperparameters as much as possible. The experiment results are shown in Figs. 4 and 5.

There is a significant difference between the training data distribution of the Big 5 and that of MBTI [7, 42] Obviously, for Big 5 personality traits, we get the best results with ALBERT-base. For MBTI personality traits, the best pre-trained language model is BERT-large. We will carry out follow-up experiments with ALBERT-base and BERT-large.

4.5 Performance comparison with multiple pre-trained language models

To verify that our framework for the Big 5 is state-of-the-art, we compare it to the following baselines, which have been reported in the past two years.

- BERT-MLP represents a MLP model using features extracted by BERT-base [7].

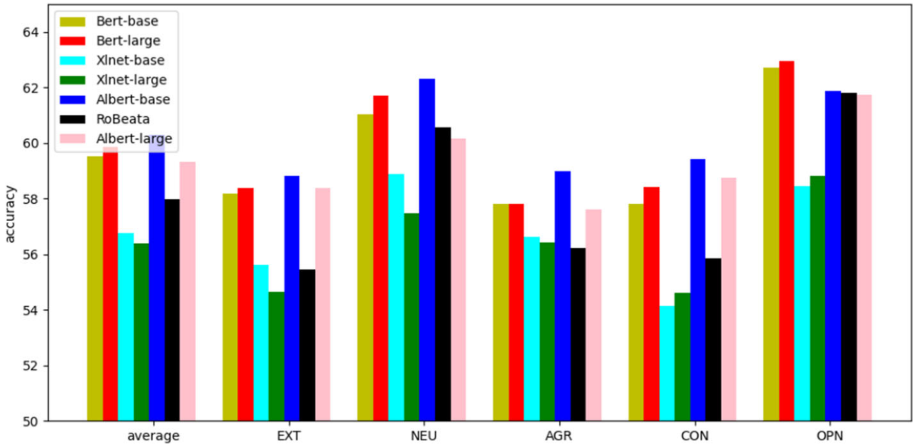

Fig. 4 Results of Performance Comparison With Multiple Pre-Trained Language Models Measured by Accuracy on Essays dataset

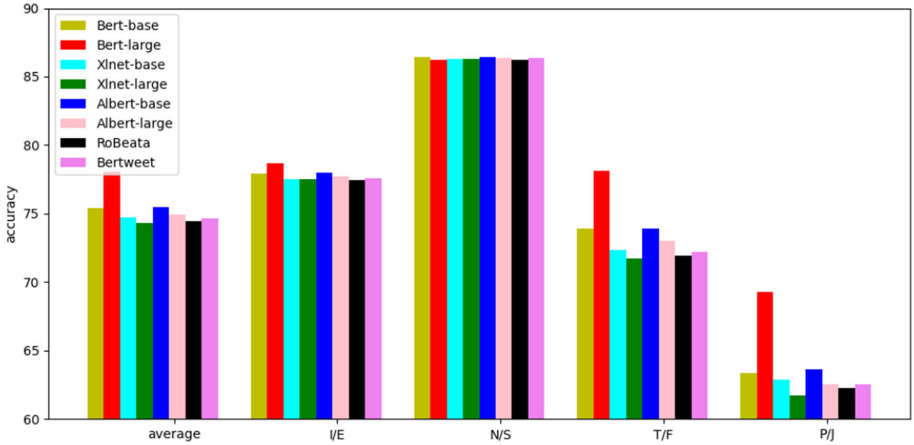

Fig. 5 Results of Performance Comparison With Multiple Pre-Trained Language Models Measured by Accuracy on Kaggle MBTI dataset

- RoBERT represents a MLP model using features extracted by RoBERT [6].
- Personality GCN represents a graph convolutional network learned from the proposed user personality graph [33].
- SEPRNN represents a MLP model only using the left and right context semantics of words extracted by BiGRU [34].
- CNN+AdaBoost represents a model with various features obtained from various filters of the convolutional neural network are fed to an AdaBoost [35].
- BERT-fusion is a model using both data and classifier level fusion. The features it adopts are extracted by three pre-trained language models [20].

In the experiments, we use a single-layer Bi-LSTM with 128 units and attention mechanism as a classifier. Other model structures and hyper parameters are the same as that in the above experiments. The adopted features total 996 dimensions. Table 4 report the accuracy and F1 of all baseline models and DLP-Personality Detection. We give preference to citing the experimental results in the above papers, and if not, we reproduce their model with the hyperparameters we optimized.

As shown in Table 4, we achieve state-of-the-art results on the Essays dataset. Specifically, the accuracy for each Big 5 personality trait achieved by our framework beat the current state-of-the-art by 28.78%, 16.39%, 7.71%, 20.37%, and 11.22%, respectively. Except for “AGR”, we achieve state-of-the-art results. Specifically, the F1 for each other Big 5 personality trait achieved by our framework beat the current state-of-the-art by 1.78%, 1.29%, 5.59%, and 10.33%, respectively.

Since there is a performance variance of our model based on the weight initialization and sample order, we report an aggregated 10-fold cross-validation performance of the outer re-sampling loop. Figures 6 and 7 report the results of cross-validation on the Essays dataset.

We achieve state-of-the-art results but the accuracy of our model has a high fluctuation, that is, it is not robust enough. The results are not enough to doubt our framework, because most of the worst value of our framework is better than the average value of current state-of-the-art model.

Table 4 Results of performance comparison on the essays dataset

Model	Big 5 (accuracy)					
	Avg	EXT	NEU	AGR	CON	OPN
BERT-MLP	60.6	60	60.5	58.8	59.2	64.6
RoBERT	–	60.62	61.07	59.72	58.55	65.86*
Personality GCN	60.92	60	63*	57.7	59.1	64.8
SEPRNN	60.26	58.48	59.71	61.20*	58.40	64.43
CNN+AdaBoost	61.87*	61.85*	62.08	59.92	64.93*	60.56
BERT-fusion	61.85	61.15	62.2	60.8	59.52	65.6
Ours	**74.07**	**79.65**	**73.33**	**65.92**	**78.16**	**73.25**
Model	Big 5 (F1)					
BERT-MLP	57.12	58	56.21	56.14	57.03	58.2
RoBERT	–	58.4	56.37	56.87	56.2	59
Personality GCN	60.92	60	63	57.7	59.1	64.8
SEPRNN	65.91	71.5*	62.36	**71.92***	63.46	67.84*
CNN+AdaBoost	68*	67	69*	69	68*	67
BERT-fusion	60.03	61.03	59.5	58.37	59.45	61.78
Ours	**70.83**	**72.77**	**69.85**	66.84	**71.8**	**74.85**

Likewise, to verify that our framework for the MBTI is state-of-the-art, we compare it to the following baselines.

- Bagged-SVM represents a model using features extracted by BERT and Mairesse features to feed to Bagged-SVM [5].
- BERT-MLP represents a MLP model using features extracted by BERT-large [7].
- RoBERT represents a MLP model using features extracted by RoBERT [6].

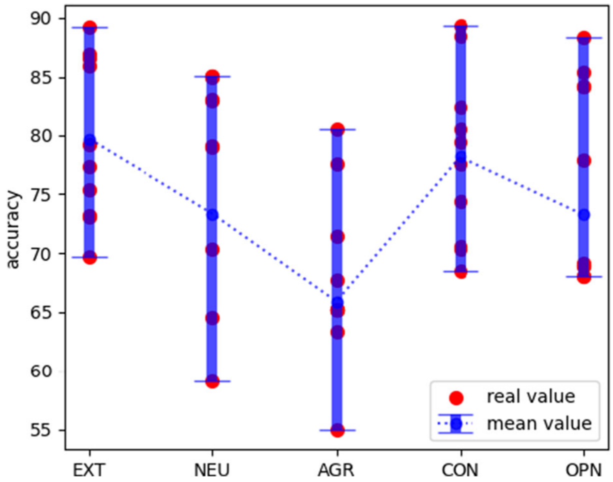

Fig. 6 Accuracy variance by 10-fold corss-validation on the Essays dataset. Where, variance of performance of “AGR” is maximal and variance of performance of “EXT” is minimal

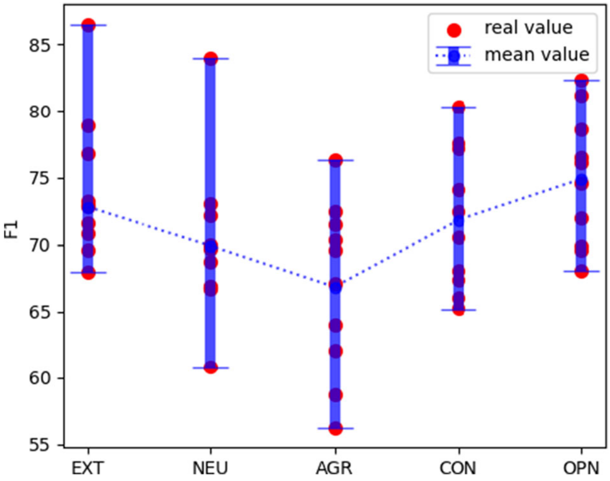

Fig. 7 F1 variance by 10-fold corss-validation on the Essays dataset. Where, variance of performance of "NEU" is maximal and variance of performance of "CON" is minimal

- LSTM+RMSprop represents a LSTM model using the RMSprop optimizer [36].
- Transformer-MD represents a Multi-Document Transformer model with a dimension attention mechanism to focus each personality dimension on the relevant post [37].
- TrigNet+GAT represents a graph network that injects structural psycholinguistic knowledge in LIWC [38].

In this experiment, the adopted features total 1147 dimensions. Table 5 report the accuracy and F1 of all baseline models and DLP-Personality Detection.

Table 5 Results of performance comparison on the Kaggle MBTI dataset

Model	MBTI (accuracy)				
	Avg	I/E	N/S	T/F	P/J
Bagged-SVM	76.1	79.0*	86.0	74.2	65.4
BERT-MLP	77.1	78.8	86.3	76.1	67.2
RoBERT	75.27	77.73	86.42	73.71	63.24
LSTM+RMSprop	75.71	77.35	86.34	72.85	66.28
Transformer-MD	77.33	76.69	86.45*	78.21*	67.98
TrigNet+GAT	77.48*	77.43	86.37	78.07	68.06*
Ours	**78.75**	**79.49**	**87.03**	**78.37**	**70.11**
Model	MBTI (F1)				
Bagged-SVM	62.72	56.67	52.85	75.42	65.94
BERT-MLP	67.31	68.05	79.35*	66.1	55.73
RoBERT	60.61	58.33	53.88	69.36	60.88
LSTM+RMSprop	63.98	61.24	67.68	68.21	58.79
Transformer-MD	70.47	66.08	69.10	**79.19***	67.50
TrigNet+GAT	70.86*	69.54*	67.17	79.06	67.69*
Ours	**78.52**	**78.61**	**87.37**	78.74	**69.32**

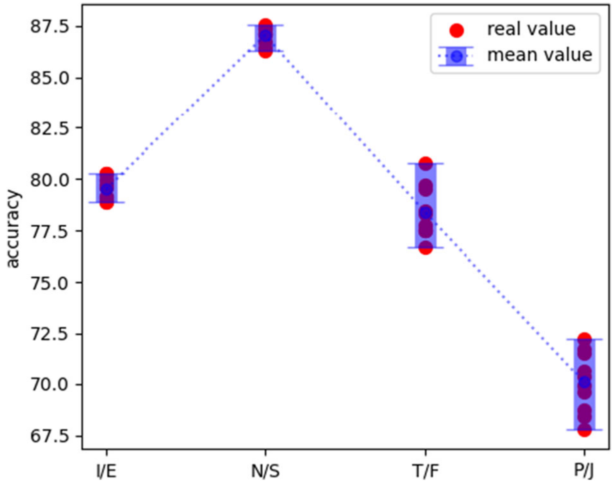

Fig. 8 Accuracy variance by 10-fold corss-validation on the Kaggle MBTI dataset. Where, the variance of performance of "P/J" is maximal and variance of performance of "N/S" is minimal

As shown in Table 5, we achieve state-of-the-art results on the Kaggle MBTI dataset. The accuracy of our framework beat the current state-of-the-art by 1.27% approximately. Specifically, the accuracy for each MBTI personality trait achieved by our framework beat the current state-of-the-art by 0.62%, 0.67%, 0.21%, and 3.01%, respectively. Except for "T/F", the F1 for each other MBTI personality trait achieved by our framework beat the current state-of-the-art by 13.04%, 10.11%, and 2.41%, respectively.

As a supplement, Figs. 8 and 9 report the results of cross-validation on the Kaggle MBTI dataset. It proves that the generalization of our model is excellent for the MBTI.

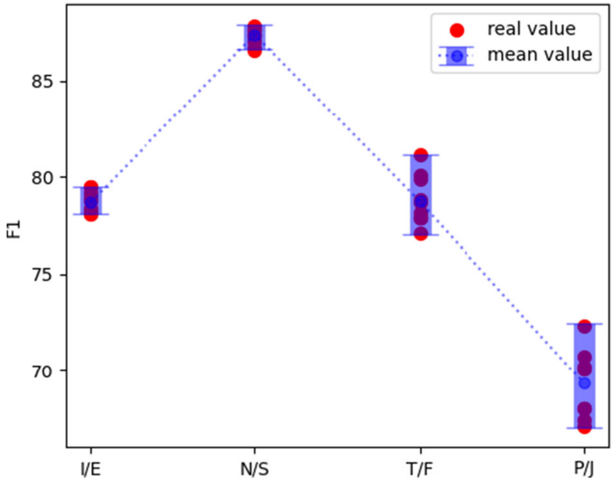

Fig. 9 F1 variance by 10-fold corss-validation on the Kaggle MBTI dataset. Where, the variance of performance of "P/J" is maximal and variance of performance of "N/S" is minimal

5 Conclusion

In this paper, we propose the DLP-Personal Detection, a novel personality detection framework, whose performance is verified by multiple experiments. As far as we know, it is the first attempt to construct a text-based personality detection framework with such high-dimensional features. It is also the first time to perform ablation experiments with such high-dimensional features. Our results show that the DLP-Personal Detection consistently beats the current state-of-the-art on the Essays and Kaggle MBTI dataset with a less complex classification network structure. Our framework obtained 74.07% accuracy and 70.83% F1 on Essays dataset. It obtained 78.75% accuracy and 78.52% F1 on Kaggle MBTI dataset. Our work will help with many potential applications, such as public opinion analysis, company management, and human-computer interaction. Importantly, it will lead personality detection research to avoid the misunderstanding of over-reliance on pre-trained language models.

The limitations of our study and future work are as follows:

- "Multimodal Learning" [39] must be the future of personality detection. The multimodal training of personality detection models with multi-source heterogeneous data such as images, audio, video, social software, and even EEG is our future work.
- Personality detection is not the end of the personality calculation. The results of personality detection should be further analyzed to make the personality detection model used in research and life [40, 41].

Data Availability The data that support the findings of this study are available from https://github.com/ml-papers-coders/Keras-BigFive-personality-traits/blog/ and https://www.kaggle.com/datasets/datanaek/mbit-type.

Declarations

Ethics Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest All authors declare that they do not have any conflict of interest.

References

1. Alessandro V, Gelareh M (2014) A survey of personality computing. IEEE Trans Affect Comput 5(3):273–291. https://doi.org/10.1109/TAFFC.2014.2330816
2. Shumanov M, Johnson L (2021) Making conversations with chatbots more personalized. Comput Hum Behav 117:106627. https://doi.org/10.1016/j.chb.2020.106627
3. Aguiar JJB, Fechine JM, Costa EB (2020) Collaborative filtering strategy for product recommendation using personality characteristics of customers. In: Proceedings of the brazilian symposium on multimedia and the web. Association for computational linguistics, pp 157–164. https://doi.org/10.1145/3428658.3430969
4. Majaluoma S, Seppala T, Kautiainen H, Korhonen P (2020) Type D personality and metabolic syndrome among Finnish female municipal workers. BMC Womens Health 20(1):202. https://doi.org/10.1186/s12905-020-01052-z
5. Kazameini A, Fatehi S, Mehta Y, Eetemadi S, Cambria B (2020) Personality trait detection using bagged svm over bert word embedding ensembles. In: The ACL 2020 workshop on Widening NLP. Association for computational linguistics
6. Jiang H, Zhang XZ, Choi DJ (2020) Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings. In: Proceedings of the AAAI

conference on artificial intelligence (Student Abstract). Association for the advancement of artificial intelligence, pp 13821–13822. https://doi.org/10.1609/aaai.v34i10.7182

7. Mehta Y, Fatehi S, Kazameini A, Stachl C, Cambria E, Eetemadi S (2020) Bottom-up and top-down: predicting personality with psychopsycholinguistic and language model features. In: Proceedings of 2020 IEEE international conference on data mining. IEEE, pp 1184–1189. https://doi.org/10.1109/ICDM50108.2020.00146
8. Zhu H, Li L, Jiang H (2018) Inferring personality traits from user liked images via weakly supervised dual convolutional network. In: The joint workshop of the 4th workshop on affective social multimedia computing and first multi-modal affective computing of large-scale multimedia data. Association for computing machinery, pp 63–69. https://doi.org/10.1145/3267935.3267953
9. Zen G, Lepri E, Ricci E, Lanz O, Bruno F, Fbkirst K (2020) Space speaks: towards socially and personality aware visual surveillance. In: ACM Int'l workshop on multimodal pervasive video analysis. Association for computing machinery, 2020, pp 37-42. https://doi.org/10.1145/1878039.1878048
10. Quercia D, Kosinski M, Stillwell D, Crowcroft J (2011) Our twitter profiles, our selves: Predicting personality with twitter. In: Proceedings of the 3rd international conference on privacy, security, risk and trust and the 3rd international conference on social computing,2011, pp 180-185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26
11. Li W, Hu X, Long X, Tang L, Chen J, Wang F, Zhang D (2020) EEG responses to emotional videos can quantitatively predict big-five personality traits. Neurocomputing 415:368–381. https://doi.org/10.1016/j.neucom.2020.07.123
12. Wang Y, Zheng J, Li Q, Wang C, Zhang H, Gong J (2021) Xlnet-caps: personality classification from textual posts. Electronics 10(11):1360. https://doi.org/10.3390/electronics10111360
13. Tausczik Y, Pennebaker J (2010) The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 29(1):24–54
14. Stajner S, Yenikent S. (2020) A survey of automatic personality detection from texts. In: Proceedings of the 28th international conference on computational linguistics. Association for computational linguistics, pp 6284-6295. https://doi.org/10.18653/v1/2020.coling-main.553
15. Mairesse F, Walker M, Mehl M, Moore R (2007) Using psycholinguistic cues for the automatic recognition of personality in conversation and text. J Artif Intell Res 30:457–500. https://doi.org/10.1613/jair.2349
16. Argamon S, Koppel DSM, Pennebaker J (2005) Lexical predictors of personality type. Proceedings of the joint annual meeting of the interface and the classification society of north america 2005:1–16
17. Nguyen T, Phung D, Adams B, Venkatesh S (2011) Towards discovery of influence and personality traits through social link prediction. In: Proceedings of the international AAAI conference on web and social media. Association for the advancement of artificial intelligence, 2011, pp 566-569. https://ojs.aaai.org/index.php/ICWSM/article/view/14151
18. Poria S, Gelbukh A, Agarwal B, Cambria E, Howard H (2013) Common sense knowledge based personality recognition from text. In: Mexican international conference on artificial intelligence. Springer, 2013, pp 484-496. https://doi.org/10.1007/978-3-642-45111-9_46
19. Vasquez RL, Ochoa-Luna J (2021) Transformer-based approaches for personality detection using the mbti model. In: XLVII latin american computing conference (CLEI). IEEE, 2021, pp 1-7. https://doi.org/10.1109/CLEI53233.2021.9640012
20. El-Demerdash K, El-Khoribi RA, Mahmoud A, Shoman I (2022) Deep learning based fusion strategies for personality prediction. Egypt Inform J 23:47–53. https://doi.org/10.1016/j.eij.2021.05.004
21. Lopez-Pabon FO, Orozco-Arroyave JR (2022) Automatic personality evaluation from transliterations of youtube vlogs using classical and state-of-the-art word embedding. Ingenierıa e Investigacion 42(2) e93803. https://doi.org/10.15446/ing.investig
22. Ren Z, Shen Q, Diao X, Xu H (2021) A sentiment-aware deep learning approach for personality detection from text. Inf Process Manag 58(3):102532. https://doi.org/10.1016/j.ipm.2021.102532
23. Jason W, Kai Z (2019) EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp 6382-6388
24. Coulombe C (2018) Text data augmentation made simple by leveraging NLP cloud APIs. arXiv:1812.04718
25. Mohammad SM (2013) Turney PD (2013) Crowdsourcing a word-emotion association lexicon. Comput Intell 29(3):436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x
26. Mohammad S (2018) Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In: Proceedings of the 56th annual meeting of the association for computational linguistics. Association for computing machinery, pp 174-184. https://doi.org/10.18653/v1/P18-1017

27. Chaturvedi I, Satapathy R, Cavallari S, Cambria E (2019) Fuzzy commonsense reasoning for multimodal sentiment analysis. Pattern Recognit Lett 125:264–270. https://doi.org/10.1016/j.patrec.2019.04.024
28. Cambria E, Poria S, Hazarika D, Kwok K (2018) Senticnet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings. In: Proceedings of the AAAI conference on artificial intelligence. Association for the advancement of artificial intelligence, pp 1795-1802. https://ojs.aaai.org/index.php/AAAI/article/view/11559
29. Qiu X, Sun T, Xu Y et al (2020) Pre-trained models for natural language processing: a survey. Sci China Technol Sci 63:1872–1897. https://doi.org/10.1007/s11431-020-1647-3
30. Yang JF, Ming XD, Wang Z (2017) Are sex effects on ethical decision-making fake or real? a meta-analysis on the contaminating role of social desirability response bias. Psychol Rep 120(1):25–48. https://doi.org/10.1177/0033294116682945
31. Ronald BL (2018) Controlling social desirability bias. Int J Mark Res 61(5):534–547. https://doi.org/10.1177/1470785318805305
32. Stajner S, Yenikent S (2021) Why Is MBTI personality detection from texts a difficult task?. In: Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume, pp 3580-3589. https://aclanthology.org/2021.eacl-main.312
33. Wang Z, Wu CH, Li QB, Yan B, Zheng KF (2020) Encoding text information with graph convolutional networks for personality recognition. Appl Sci 10:4081. https://doi.org/10.3390/app10124081
34. Xue X, Feng J, Sun X (2021) Semantic-enhanced sequential modeling for personality trait recognition from texts. Appl Intell 51(11):7705–7717. https://doi.org/10.1007/s10489-021-02277-7
35. Mohades Deilami F, Sadr H, Tarkhan M (2022) Contextualized multidimensional personality recognition using combination of deep neural network and ensemble learning. Neural Process Lett. https://doi.org/10.1007/s11063-022-10787-9
36. Mawadatul M, Hilman FP (2021) Prediction of myers-briggs type indicator personality using long short-term memory. Jurnal Elektronika dan Telekomunikasi 21(2) 104-111. https://doi.org/10.14203/jet.v21.104-111
37. Yang F, Quan X, Yang Y, Yu JX (2021) Multi-document transformer for personality detection. In: Proceedings of the AAAI conference on artificial intelligence. vol 35, no 16, pp 14221-14229. https://ojs.aaai.org/index.php/AAAI/article/view/17673
38. Yang T, Yang F, Ouyang H, Quan XJ (2021) Psycholinguistic tripartite graph network for personality detection. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, pp 4229-4239. https://aclanthology.org/2021.acl-long.326
39. Huang Y, Du C, Xue Z, Xuan YC, Zhao H, Huang LB (2021) What makes multi-modal learning better than single (Provably). In: The 35th conference on neural information processing systems (NeurIPS)
40. Amitabha A, Aman A, Sujay S, Anupam G (2022) Impact of COVID-19 on the human personality: an analysis based on document modeling using machine learning tools. Comput J, bxab207
41. Shappie AT, Dawson CA, Debb SM (2020) Personality as a predictor of cybersecurity behavior. Psychol Pop Media 9(4):475–480
42. Fabio C, Lepri B (2018) Is big five better than MBTI? a personality computing challenge using twitter data. In: Fifth italian conference on computational linguistics