A New Contrastive Learning Based Model for Estimating Degree of Multiple Personality Traits Using Social Media Posts

Kunal Biswas[1], Palaiahnakote Shivakumara[2(✉)], Umapada Pal[3], and Ram Sarkar[1]

[1] Jadavpur University, Kolkata, India
ram.sarkar@jadavpuruniversity.in
[2] Faculty of Computer Science and Information Technology, University of Malaya, Kula Lumpur, Malaysia
shiva@um.edu.my
[3] Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
umapada@isical.ac.in

Abstract. Estimating the degree of multiple personality traits in a single image is challenging due to the presence of multiple people, occlusion, poor quality etc. Unlike existing methods which focus on the classification of a single personality using images, this work focuses on estimating different personality traits using a single image. We believe that when the image contains multiple persons and modalities, one can expect multiple emotions and expressions. This work separates given input images into different faces of people, recognized text, meta-text and background information using face segmentation, text recognition and scene detection techniques. Contrastive learning is explored to extract features from each segmented region based on clustering. The proposed work fuses textual and visual features extracted from the image for estimating the degree of multiple personality traits. Experimental results on our benchmark datasets show that the proposed model is effective and outperforms the existing methods.

Keywords: Scene text · Contrastive Learning · Text appearance · Sentiment analysis · Image personality traits

1 Introduction

Personality traits identification is a vital issue in the everyday life of the person such as the results of elections and verdicts of courts [1]. In addition, it can also be used to find leader, assess student performance, interview and selection of students for research or jobs or defense [2]. For all the above applications, it is necessary to estimate the degree of different personality traits in the images. This is because multiple personality traits, like multiple skills, such as agreeableness, openness, and conscientiousness are required to choose suitable candidates especially for hiring software employees and research. At the same time, a few applications may look for particular personality traits irrespective of the degree of personality traits like finance, management, security jobs etc. It is also true

H. Lu et al. (Eds.): ACPR 2023, LNCS 14407, pp. 15–29, 2023.
https://doi.org/10.1007/978-3-031-47637-2_2

that if we look at the facial or posture, pose and background information of the person in the case of images and multiple words in the text lines, one can observe multiple personality traits rather than a single personality trait.

It is illustrated in Fig. 1, where we can see for each image, different personality traits according to textual, facial and background information. For example, (i) Openness: From the image, we can see a creative mask being worn for an ID photo which indicates the openness to experience trait. Also, from the text, the word "*new*" is recognized, which also indicates the openness class, which is the dominant trait. Also, from the dark imagery and the intention of hiding the face, we can infer high traits of neuroticism from the image. (ii) Conscientiousness: The text recognized contains the word "*educate*" which is an indicator of class conscientiousness, which is the dominant trait. We can also infer the trait of agreeableness due to the inference of the words "*reduce poverty*". (iii) Extraversion: The image shows a huge gathering of people all with happy faces and enjoying themselves. This is highly indicative of the trait of extraversion. We can also see that none of the other traits is high in this image. (iv) Agreeableness: The image shows a sympathetic person and the text recognized contains the word "*sorry*" which are both indicative of the trait of agreeableness, which is dominant here. (v) Neuroticism: The image shows a man hiding in dark clothing and the text recognized contains "*madness*". Also, the text infers that a person must not lose their madness which is very indicative of mental instability and the trait neuroticism, which is dominant here.

Therefore, one can infer that for better solutions and decisions, assessing and estimating the degree of multiple personality traits from a single image is vital for many day-to-day situations. For personality traits identification, several methods have been developed in the past using facial information [1], handwritten text information [3], normal text uploaded on social media information [4], and profile information, which provides multimodal information [4]. Most methods aim at detecting single personality traits for the input image or text information. Similarly, if personality traits identification is considered as a classification problem, the methods developed for classification may not perform well for the images containing person, actions and postures etc. [5]. This is due to these models are limited to scene images. This observation has motivated us to introduce multiple personality traits identification in a single image through this work. To the best of our knowledge, this is the first of its kind work.

To localize the content in the image, the proposed work detects text in the images and segments words from the text line. We believe the meaning of text directly represents personality traits like facial information. Therefore, for each word, the proposed method obtains feature vectors. In addition, for the whole input image, the proposed work uses Google Cloud Vision API [6] to obtain labels (text annotations). Besides, if any text, like, the description of the image and profile picture, the proposed work obtains a feature vector. The feature vectors obtained from each word, annotations and description of the images are considered as textual features. To strengthen the feature extraction for personality traits identification, the proposed work segments the whole image into face and background clusters (other than the face and text information in the input image). The features extracted from the image content are called visual features. Finally, the model fuses textual and visual features for estimating multiple personality traits in a single image. Feature extraction from textual and image information is motivated by

contrastive learning [7], which has special properties for representing visual and textual features. Then classification is done using an Artificial Neural Network (ANN) model by feeding feature vectors as inputs.

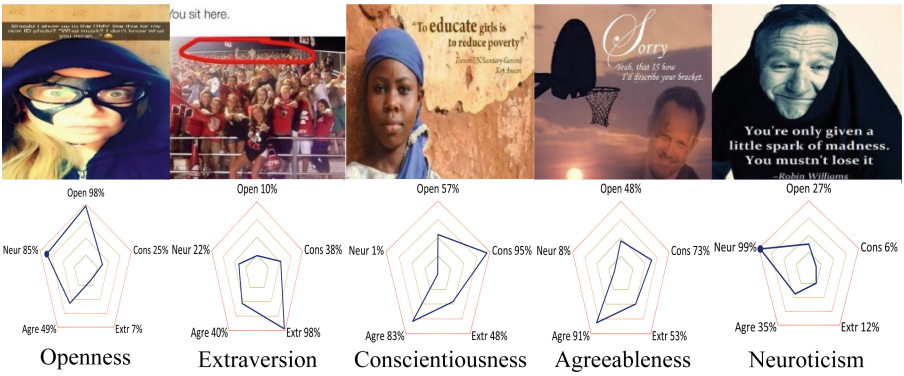


Fig. 1. Sample images with different degrees of personality traits assessments.

The main contributions of the work are as follows: (i) Proposing a new method to estimate multiple personality traits from a single image. (ii) Exploring contrastive learning for feature extraction from visual and textual features is novel here. (iii) The weighted approach for fusing visual and textual features for multiple personality traits estimation is also a novel contribution.

2 Related Work

For personality traits identification/assessment, in the past, many methods have been proposed based on normal text uploaded on social media, handwritten text with a graphological approach, facial information and image-text information. We review some of the methods in this section.

Kumar et al. [8] proposed a language embedding-based model for personality traits assessment using social network activities. The work considers the review or any text which describes personality traits uploaded on social media for personality traits assessment. Anglekar et al. [9] proposed a method based on deep learning for self-assessment, especially for interview preparations. Dickmond et al. [10] explored machine learning for extracting features from curriculum vitae to identify personality traits. It is noted from the above methods that the scope of the models is limited to textual information, and researchers ignored image information for personality traits identification. Therefore, these methods may not be effective for estimating the degree of different personality traits in a single image.

There are methods for personality trait identification using a graphology-based approach, which use handwritten characters to study the behavior of the writer [3, 11, 12]. Although, the models used image information, developed based on an unscientific

approach. Therefore, the results of the modes may not be consistent and reliable for estimating the degree of different personality traits in the images.

To improve the performance of personality trait identification, some models used image and textual information [3, 13, 14]. For example, Sun et al. [2] developed a method to evaluate the aptitude and entrance psychological aspects of the students. The approach extracts features from the facial region for assessing personality traits. Beyan et al. [15] extracted non-verbal features from key dynamic images for personality trait identification. The combination of convolutional and long short-term memory models has been proposed for improving performance of the personality traits identification. Xu et al. [16] proposed a model for predicting five personality traits using static facial images and different academic backgrounds. For a given input video frame, Ventura et al. [1] estimated the degree of different personality traits. This model requires a still frame, multiple frames and audio to achieve the best results. Therefore, this model may not perform well for a single image.

Since sentiment analysis is related to personality traits identification, we review the methods of sentiment analysis to show that the methods are not effective for estimating degree of personality traits. For instance, Yu et al. [17] proposed a transformer-based based step to fuse image and text modalities with self-attention layers for sentiment analysis. Thus, estimating the degree of multiple personality traits for a given input image remains challenging. Hence, this work focuses on developing a new model based on segmented local information and contrastive learning for assessing multiple personality traits in the image.

3 Proposed Model

For a given input image, the proposed method segments facial and textual regions and considers other than textual and facial information as background. It is true that each segment provides cues of different personality traits if the image contains multiple people with emotions, expressions and actions. Therefore, this work segments the text using Google Cloud Vision API, face using the Haar cascades algorithm [18]. For the background, the method uses the Google Cloud Vision API for obtaining labels. In the same way, for the detected text and description of the image/profile, this work uses Google Cloud Vision API for obtaining recognition results.

Overall, the Google Cloud Vision API provides recognition results for the text in the image, description of the image, and background scene of the image, which are grouped as textual information, and facial information as visual information. For extracting features from textual and facial information, inspired by the success of contrastive learning for discriminating the objects based on similarity and dissimilarity [6], we explore the same for feature extraction process in this work. The extracted features are fused for estimating multiple degrees of personality traits in the image through five individual neural network regression models. The pipeline of the proposed method can be seen in Fig. 2. For a given input image, the work segments parts for feature extraction are represented as defined in Eq. (1).

$$I_S = T_R + T_A + T_B + I_F \tag{1}$$

where, I_S stands for sample image, T_R stands for text recognized, T_A stands for annotated text, T_B stands for background scene text and I_F stands for facial part. The face is detected using the Haar Cascades algorithm [18], which works based on the sum of the pixel intensities in each region and the differences between the sums. The process is performed in a cascading fashion. The extracted features are then supplied to a machine-learning algorithm for face detection. The detected face and text information is removed from the input image and replaced with a black rectangular patch. This results in a background image as defined in Eq. (2).

$$I_B = I_S - I_F \tag{2}$$

where, I_B is the background image. The background image is fed to the recognizer (Google Cloud Vision API) to determine labels.

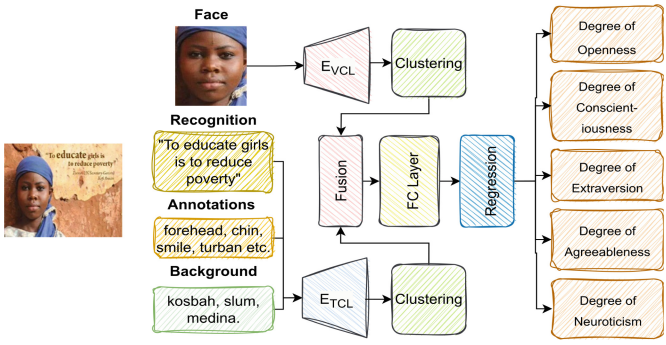


Fig. 2. Block diagram of the proposed method used to estimate the degree of multiple personality traits. The face, text recognition, annotation and background are extracted from the sample image and encoded. Then clustering and fusion steps followed by regression are done to estimate the degree of personalities.

3.1 Contrastive Learning for Textual Feature Extraction

This work considers three types of texts for each sample. Firstly, text recognized from the image is obtained using Google Cloud Vision API [6]. Secondly, image annotations from the sample image are obtained again using [6]. Only the top ten most confident annotations are taken into consideration. Thirdly, background scenes are recognized using ResNet18 [19] which has been trained on the Places dataset [20]. This dataset has 10 million images comprising 400+ unique scene categories. From this model, we choose the top 10 highest confidence scene categories as the background scenes. While assessing the personality traits, the personality trait with the highest degree is known as the dominant personality trait. All texts obtained from the previous step are compiled into a corpus with the text parts as individual documents and the dominant personality trait as the document label.

The text within the corpus is then used to train the representation space. This is done using contrastive learning. A pair of text from the same class labels is taken and fed

through a neural network to get a pair of outputs. The similarity score between the pair of outputs from the neural network is calculated using Eq. (3). In another instance a pair of text from two different classes is taken and fed through the same neural network. Again, the similarity score between these outputs is calculated using Eq. (3). Finally, the loss is calculated using Eq. (4). This loss is propagated backwards to the network learns to optimize the loss. This output space of the learned neural network is known as the textual representation space. A pair of texts with different labels repel each other as they are contrastive samples, whereas similar labels attract each other. This is done using NT-Xent loss. This loss uses a similarity function.

$$sim(t_a, t_b) = \frac{{t_a}^T t_b}{\|t_a\| \|t_b\|} \tag{3}$$

where t_a represents text part a and t_b represents textual part b, t_a^T is the transpose of t_a and $\|t_a\|$ represents the norm of t_a. NT-Xent Loss [6] is given by:

$$L_{a,b} = -\log \frac{e^{sim(t_a, t_b)/\tau}}{\sum_{k=1}^{2N} 1_{[k \neq i]} e^{sim(t_a, t_b)/\tau}} \tag{4}$$

where, $1_{[k \neq i]}$ is an indicator function evaluating to 1 iff $k \neq i$, and $sim(t_a, t_b)$ comes from Eq. (3). Hence minimizing this function in Eq. (4) over all pairs of attractive and repulsive pairs gives us a text representation space. Since there is a variable number of words, and a variable dimensional input to a neural network is not possible, there is a need to fix the number of dimensions. This is performed by clustering the words in the representation space into 10 clusters using K-means clustering. The clusters centroids from the representation space are then considered as the text embeddings. Separate text embeddings for text recognized annotations and background scenes each with 10 centroids are obtained. The value of 10 clusters is determined empirically. This process has been shown in Fig. 3.

$$E_R = TCLR(T_R), E_A = TCLR(T_A), E_B = TCLR(T_B) \tag{5}$$

where, T_R, T_A, and T_B are taken from Eq. (1), and $TCLR()$ is the Textual Contrastive Learning Representation function.

3.2 Contrastive Learning for Visual Feature Extraction

The sample image may contain many facial regions owing to multiple people in the same image. Motivation has been drawn from other visual learning models like [21]. These facial images are labelled according to their dominant personality traits. Now, the visual representation space for facial images is trained using contrastive learning in the same way. Similar to the textual counterpart, two facial images belonging to the same class are considered and supplied to a neural network. The similarity between the outputs is considered and loss is calculated. This loss is propagated back through the network to bring the facial images of same class closer. Facial images of varying classes are used further away with a pair of facial images from different classes. This results in learning visual representation space. Hence this framework can be used for both texts

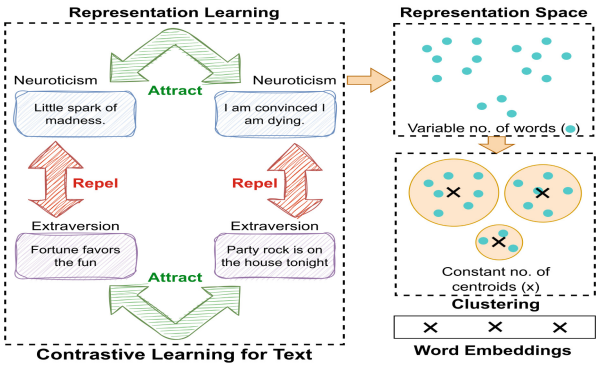


Fig. 3. Block diagram for word embeddings using contrastive learning. The recognized words from same class attract each other whereas, the recognized words from differing class repel each other.

as well as images. The same NT-Xent loss is used for training as shown in Eq. (4). This results in an image embedding for every facial image in the visual representation space. Again, there can be multiple people in a single image resulting in multiple facial image embeddings. To supply this to a neural network, the length of the embedding needs to be fixed. Again, clustering is used in the visual representation space on the facial images. K-means clustering has been used to find two cluster centroids. These are the facial image embeddings. The method has been shown in Fig. 4.

$$E_F = VCLR(I_F) \tag{6}$$

where, I_F is taken from Eq. (1) and $VCLR()$ is the Visual Contrastive Learning Representation function.

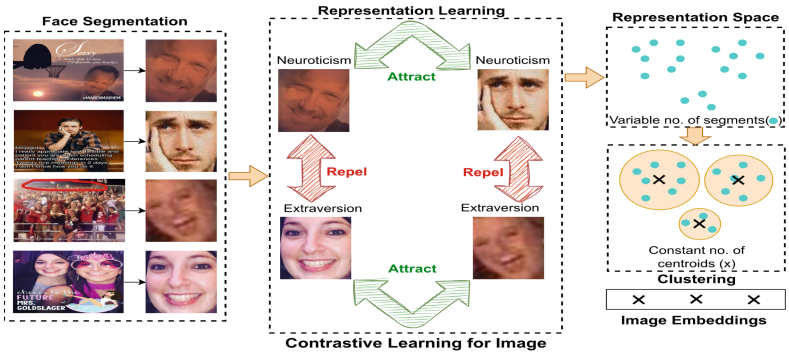


Fig. 4. Block diagram for image embeddings using contrastive learning. The recognized facial images from same class attract each other whereas, the recognized facial images from differing class repel each other.

3.3 Weighted Fusion and Estimating Degree of Personality Traits

In this step, the embeddings obtained in the previous step are fused to obtain the final feature vector F_v. This is performed using a weighted approach. The objective here is to assign the weights so that the features which are more discriminating are given priority. With this motivation, variance has been used as a metric that can evaluate the discriminating power of a feature. The embeddings of text recognized, annotation, background, and facial image are concatenated, and the variance is calculated. The embeddings of all the parts, other than the text recognized, is concatenated and variance is calculated, for calculating for the weightage of text recognized. Then the difference between these two variances is taken and the square root is the weightage of the text recognized and represented by E_R. This calculation has been shown in Eq. (7).

$$W_R = \sqrt{\left[\sigma^2(E_R \oplus E_A \oplus E_B \oplus I_F) - \sigma^2(E_A \oplus E_B \oplus I_F)\right]^2} \tag{7}$$

where, σ^2 is the variance, $\oplus$ stands for concatenation, E_R, E_A, and E_B come from Eq. (5), I_F comes from Eq. (6), and W_R is the weightage of text recognized. Similarly, the weightage of annotations, background scene text and facial image are calculated as follows:

$$W_A = \sqrt{\left[\sigma^2(E_R \oplus E_A \oplus E_B \oplus I_F) - \sigma^2(E_R \oplus E_B \oplus I_F)\right]^2} \tag{8}$$

where, W_A stands for the weightage of annotation. Similarly, W_B, and W_F, which are the weights of background and face respectively, are obtained using Eq. (8). The weightage of text recognition, W_R, is multiplied with the text recognized embedding, E_R to obtain the final weighted text recognition embedding. Similarly, weightage of annotation, W_A, is multiplied with annotation embedding, E_A, weightage of annotation background, W_B, is multiplied with background embedding, E_B, and weightage of facial images, W_F, is multiplied with facial embedding, E_F to obtain the respective final embeddings. These four embeddings are concatenated to obtain the final feature vector as shown in Eq. (9).

$$F_v = W_R \cdot E_R \oplus W_A \cdot E_A \oplus W_B \cdot E_B \oplus W_F \cdot E_F \tag{9}$$

where, F_v is the final feature vector. This process has been shown in Fig. 5.

The final feature vector F_v is supplied to five separate neural networks regression models as shown in Fig. 5 for classification. Each network has four hidden layers with 512, 256, 64, and 16 activation units with the rectified linear unit (ReLU) activation function [22]. The output layer has one unit. The loss function used is Mean Squared Error (MSE) [23] with Adam optimizer [24]. The batch size has been set to 32 with learning rates of 0.001. The training is run for 50 epochs for each neural network separately to obtain the predictions, as shown in Fig. 5.

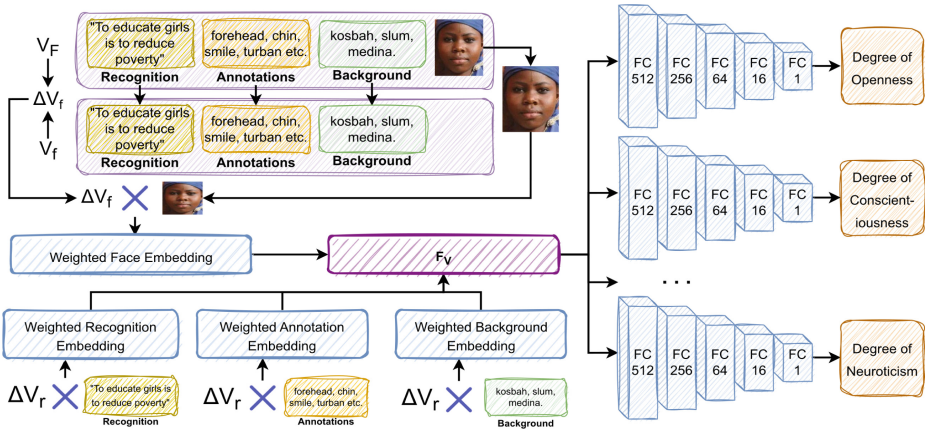


Fig. 5. Weighted fusion technique based on Variance and Estimating degree of personality traits. The representations are multiplied with their weightage and concatenated to form the final feature vector F_v. This F_v is then supplied to FC layers for estimating degrees of personality.

4 Experimental Results

Our dataset includes 5000 Twitter images that are collected from the source given in [25], which provides 559 Twitter user accounts and image posts. For all the collected images, the ground truth is generated via a questionnaire NEO-PI-R [26], which is a standard procedure used for psychological data collection. In the test, responses from the users are recorded for a given multiple questions, and the responses are measured on a Likert scale from 1–5. The weightage has been assigned according to the question. The average of weights is considered the actual label of a personality trait. This calculation results in fine-grained personality trait scores for each individual personality per image. To test the performance of the proposed method on a standard dataset, the PERS Twitter dataset [27] which is a considerably large dataset containing 28434 profile pictures, has been used.

To test the effectiveness of the proposed method, a comparison is done with the following state-of-the-art models. Biswas et al. [4], use a multimodal concept for personality traits identification using Twitter posts. Wu et al. [5], use vision transformers for the classification of images. Further, we also implemented the state-of-the-art method [17] developed for sentiment analysis to show that sentiment analysis method is ineffective for estimating degree of personality traits in the images. The existing methods are fine-tuned with the training sample as the proposed method for experimentation. The same setup is used for all the experiments. The ratio of 70:30 for training and testing is considered for all the experimentation.

For measuring the performance of the method, we use three standard metrics that are (i) Spearman's Correlation Coefficient, (ii) Pearson's correlation, and (iii) Root Mean Square Error (RMSE). Spearman's Correlation Coefficient: Spearman's correlation coefficient is a nonparametric measure of the rank correlation between two variables as defined in Eq. (10). A high value indicates better performance.

$$R_S = 1 - \frac{6\sum_{i=1}^{n}(t_i - p_i)}{n(n^2-1)} \tag{10}$$

where, t_i is the ground truth value, p_i is the predicted value and n is the number of samples. Pearson's correlation coefficient: Pearson's correlation coefficient is the test statistic that measures the statistical relationship, or association, between two continuous variables. It is defined in Eq. (11). A high value indicates better performance.

$$R_P = \frac{n\sum t_i p_i - \sum t_i \sum p_i}{\sqrt{\left[n\sum t_i^2 - (\sum t_i)^2\right]\left[n\sum p_i^2 - (\sum p_i)^2\right]}} \tag{11}$$

where, n is the number of samples. Root Mean Square Error (RMSE): RMSE shows how far predictions fall from measured true values using the Euclidean distance. It is defined in Eq. (12). A low value indicates better performance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\|t_i - p_i\|^2}{n}} \tag{12}$$

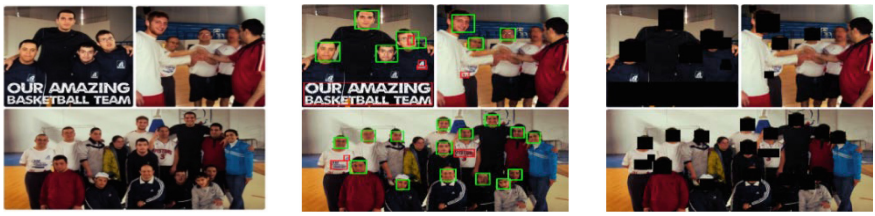
where n is the number of samples, t_i is the ground truth value, and p_i is the predicted value.

4.1 Ablation Study

To estimate the degree of multiple personality traits for each image, features, namely, textual, image and weighted fusion are the key components in the proposed model. Therefore, to validate the effectiveness of each component, we have conducted the following experiments. We use (i) Only image features, (ii) Only textual features, (iii) Image along with textual features, (iv) Image features with weighted fusion, and (v) Textual features with weighted fusion, and (vi) Image and textual features with weighted fusion for calculating measures. Table 1 shows that the results of textual features alone are better than image features alone. This shows that textual features provide more rich semantic information than image features. The same conclusions can be drawn from the experimental results shown in (iv) and (v). In the same way, when we combine image and textual features, as reported in (iii), without weighted fusion, which is better than individual features and image features with the weighted fusion. This shows that image and textual features contribute equally to achieving the best results. However, when we combined all images, textual and weighted fusion, which is actually the proposed method, reports the highest results compared to all other experiments. Therefore, one can infer that the key features are effective in coping up with the challenges of multiple personality trait estimations. Note that a " $\blacklozenge$ " in the case of weighted fusion indicates a normal concatenation, whereas, in the case of the features it indicates that the features are omitted. The reported correlations and RMSE are averaged for all the five personality traits scores.

Table 1. Assessing the contribution of each modality using our dataset.

Exp.	Image Features	Textual Features	Weighted Fusion	Rs	Rp	RMSE
(i)	✓	◆	◆	0.19	0.21	1.73
(ii)	◆	✓	◆	0.35	0.42	1.31
(iii)	✓	✓	◆	0.60	0.68	1.23
(iv)	✓	◆	✓	0.20	0.19	1.69
(v)	◆	✓	✓	0.73	0.75	1.18
(vi)	✓	✓	✓	0.81	0.86	1.12



Input images Text and face Background scene

Fig. 6. Illustrating segmentation results of the proposed method. Texts are separated by fixing bounding boxes, faces are separated by face detection step and other than a face and text is considered as the background scene.

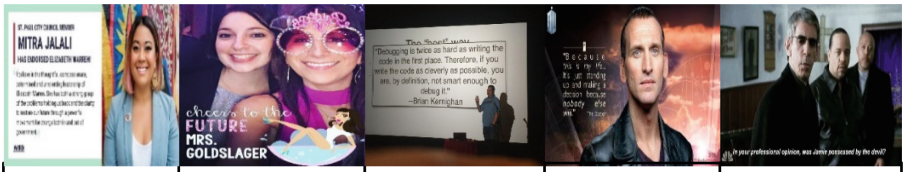
4.2 Experiments on Segmentation

For evaluating the segmentation step, qualitative results of the proposed method on sample images of different personality traits are shown in Fig. 6, where it is noted that for all three input images, the steps segment text, face and background well. Therefore, we can conclude that the steps are robust to degradation caused by social media.

4.3 Estimating Multiple Personality Traits

Qualitative results of the proposed method for estimating five personality traits for each image are shown in Fig. 7, where it is noted that for all the images of (O, C, E, A, N), the predicted values are almost close to the ground truth. This indicates that our method is capable of estimating multiple personality traits for each image. It is also noted from Fig. 7 that the predicted bold values are still close to the ground truth compared to the score of other personality traits. The dominant score can be used for the classification of single personality traits of each image instead of multiple personality traits. Therefore, the method can be used for estimating multiple personality traits as well as single personality traits. To verify the same conclusions, quantitative results of the proposed and existing methods [4, 5] on our dataset and standard dataset (PERS Twitter) are reported in Table 2 and Table 3, respectively.

The scores of Spearman's and Person's correlation coefficients, and RMSE reported in Table 2 and Table 3 on our and PERS datasets show that our method is better than existing methods in terms of all three measures except the method [4] for one personality



Pers	GT	Pre
O	**97**	**98**
C	42	37
E	59	66
A	6	2
N	60	68

Pers	GT	Pre
O	19	23
C	60	52
E	**98**	**95**
A	56	62
N	8	24

Pers	GT	Pre
O	66	58
C	**98**	**97**
E	51	58
A	81	76
N	7	14

Pers	GT	Pre
O	65	68
C	46	53
E	73	59
A	**97**	**89**
N	15	11

Pers	GT	Pre
O	62	64
C	19	12
E	45	62
A	5	13
N	**99**	**91**

Fig. 7. Example of the successful result of the proposed method. The bold represents the actual personality traits class if we consider the dominant score as the actual label.

trait. Therefore, it can be concluded that the method estimates multiple personality traits accurately for each image in terms of statistical relationship, monotonic relationship and error. Since the RMSE of the proposed method is lower than the existing methods for almost all personality traits, the predicted score is close to the ground truth. This observation confirms that the proposed method is accurate for multiple personality trait estimations. However, the existing methods report poor results because the method [4] was developed for single personality traits identification, while the method [5] was developed for image classification but not personality traits identification. In the same way, since the method [17] was developed for sentiment analysis, it does not perform well for classification of multiple personality traits in the images. For some personality traits, the method [4] achieves the highest result compared to the proposed and other existing methods [5]. This is because the objective of the method was to estimate single personality traits while the proposed method is to estimate multiple personality traits.

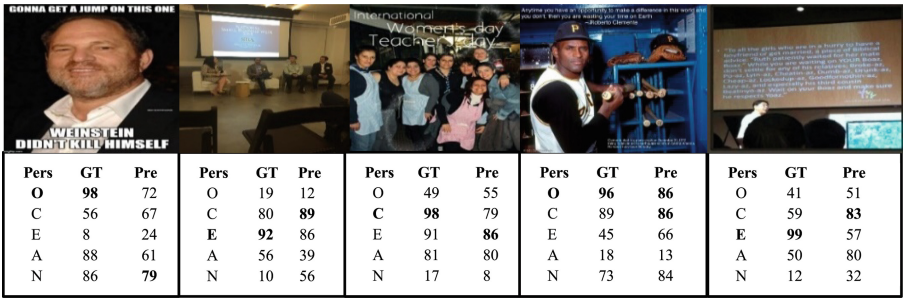
Table 2. Spearman's Correlation Coefficient, Pearson's Correlation Coefficient and RMSE for the proposed and existing methods on our dataset.

Trait	Rp				Rs				RMSE			
	Our	[4]	[27]	[17]	Our	[4]	[5]	[17]	Our	[4]	[5]	[17]
O	**0.68**	0.66	0.34	0.52	0.72	**0.78**	0.48	0.39	**1.14**	1.48	1.74	1.86
C	**0.95**	0.78	0.36	0.43	**0.98**	0.51	0.64	0.44	**1.07**	1.53	1.81	1.59
E	**0.93**	0.81	0.44	0.54	**0.86**	0.81	0.53	0.72	**1.18**	1.71	2.27	2.06
A	**0.65**	0.51	0.28	0.32	**0.98**	0.89	0.49	0.50	**1.16**	1.20	1.97	2.11
N	**0.85**	0.67	0.34	0.46	**0.78**	0.67	0.58	0.63	**1.05**	1.38	1.42	1.55

Sometimes, when the image does not provide sufficient clues for multiple personality traits estimation as shown in sample images in Fig. 8, the method fails to achieve the best results. For example, the third and fifth images from the left-top, where we can see the images lost visual features to define it as an Extraversion (ground truth). The

Table 3. Spearman's Correlation Coefficient, Pearson's Correlation Coefficient and RMSE for the proposed and existing methods on PERS Twitter Dataset.

Trait	Rp				Rs				RMSE			
	Our	[4]	[5]	[17]	Our	[4]	[5]	[17]	Our	[4]	[5]	[17]
O	0.56	**0.62**	0.24	0.31	**0.65**	0.62	0.33	0.36	**1.23**	1.41	1.58	1.59
C	**0.88**	0.64	0.31	0.59	**0.82**	0.59	0.37	0.67	**1.24**	1.38	1.77	1.45
E	**0.65**	0.51	0.28	0.32	**0.72**	0.51	0.27	0.46	**1.18**	1.51	1.61	1.67
A	**0.81**	0.69	0.25	0.21	**0.83**	0.68	0.46	0.24	**1.12**	1.26	1.42	2.16
N	**0.68**	0.54	0.19	0.35	**0.78**	0.65	0.41	0.42	1.28	**1.24**	1.54	1.39



Pers	GT	Pre
O	**98**	72
C	56	67
E	8	24
A	88	61
N	86	**79**

Pers	GT	Pre
O	19	12
C	80	**89**
E	**92**	86
A	56	39
N	10	56

Pers	GT	Pre
O	49	55
C	**98**	79
E	91	**86**
A	81	80
N	17	8

Pers	GT	Pre
O	**96**	**86**
C	89	**86**
E	45	66
A	18	13
N	73	84

Pers	GT	Pre
O	41	51
C	59	**83**
E	**99**	57
A	50	80
N	12	32

Fig. 8. Samples where poor results are reported. The bold represents the actual personality traits class if we consider the dominant score as the actual label.

method misclassifies both images as consciousness. In the same way, for the first image, the visual feature indicates Openness, while the textual feature indicates neuroticism and the ground truth indicates Openness. When there is a conflict among the features, the method fails to estimate personality traits correctly. The same conflict can be seen in the second and fourth images, and hence, the method classifies it as Neuroticism. To overcome this problem, we plan to develop an end-to-end transformer, which can cope up with the challenges of degradation and extract minute information. Therefore, there is scope for improvement in the near future.

5 Conclusion and Future Work

In this work, we have proposed a new method for estimating multiple personality traits in a single image rather than a single personality for each image. To extract the local information to study the multiple personality traits, the proposed work segments text, face and background. For each segment, we apply contrastive learning for feature extraction, which extracts textual from text information and visual features from image information. The work also introduces a weighted fusion operation for fusing textual and visual features. The features are fed to five ANN based regression models for estimating five personality traits. Experimental results on our and standard datasets show that the

proposed method outperforms the existing methods in terms of statistical relationship, monotonic relationship and mean square error. However, when the image loses clues and conflicts between the visual and textual features, the method performs poorly. We plan to deal with this in the future.

Acknowledgement. The work was supported by Ministry of Higher Education Malaysia via Fundamental Research Grant Scheme with Grant no: FRGS/1/2020/ICT02/UM/02/4. And also, this work was partially supported by Technology Innovation Hub (TIH), Indian Statistical Institute, Kolkata, India.

References

1. Ventura, C., Masip, D., Lapedriza, A.: Interpreting CNN models for apparent personality trait regression. In: Proceedings CVPRW, pp. 55–63 (2017)
2. Sun, X., Huang, J., Zheng, S., Rao, X., Wang, M.: Personality assessment based on multimodal attention network learning with category-based mean square error. IEEE Trans. Image Process. **31**, 2162–2174 (2022)
3. Alamsyah, D., Widhiarsho, W., Hasan, S., et al.: Handwriting analysis for personality trait features identification using CNN. In: Proceedings ICoDSA, pp. 232–238 (2022)
4. Biswas, K., Shivakumara, P., Pal, U., Chakraborti, T., Lu, T., Ayub, M.N.B.: Fuzzy and genetic algorithm-based approach for classification of personality traits oriented social media images. Knowl.-Based Syst. **241**, 108024 (2022)
5. Wu, H., et al.: CVT: introducing convolutions to vision transformers. In: Proceedings ICCV, pp. 22–31 (2021)
6. Google Cloud Vision AI. https://cloud.google.com/vision. Accessed 25 June 2021
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings ICML, pp. 1597–1607 (2020)
8. Kumar, P.K.N., Gavriova, M.L.: Latent personality traits assessment from social network activity using contextual language embedding. IEEE Trans. Comput. Soc. Syst. **9**(2), 38–649 (2021)
9. Anglekar, S., Chaudhari, U., Chitanvis, A., Shankarmani, R.: A deep learning based self-assessment tool for personality traits and interview preparations. In: Proceedings ICCICT (2021)
10. Dickmond, L., Hameed, V.A., Rana, M.E.: A study of machine learning based approaches to extract personality information from curriculum vitae. In: Proceedings DeSE (2021)
11. Kulsoom, S., Latif, S., Saba, T., Latif, R.: Students' personality assessment using deep learning from university admission statement of purpose. In: Proceedings CDMA (2022)
12. Gahmousse, A., Gattal, A., Djeddi, C., Siddiqi, I.: Handwriting based personality identification using textural features. In: Proceedings ICDABI (2020)
13. Biswas, K., Shivakumara, P., Pal, U., Lu, T., Blumenstein, M., Lladós, J.: Classification of aesthetic natural scene images using statistical and semantic features. Multimedia Tools Appl., 1–26 (2022)
14. Biswas, K., Shivakumara, P., Pal, U., Lu, T.: A new ontology-based multimodal classification system for social media images of personality traits. SIViP **17**(2), 543–551 (2023)
15. Beyan, C., Zunino, A., Shahid, M., Murino, V.: Personality traits classification using deep visual activity-based nonverbal features of key-dynamic images. IEEE Trans. Affect. Comput. **12**(4), 1084–1099 (2019)

16. Xu, J., Tian, W., Lv, G., Liu, S., Fan, Y.: Prediction of the big five personality traits using static facial images of college students with different academic backgrounds. IEEE Access **9**, 76822–76832 (2021)
17. Yu, J., Kai, C., Rui, X.: Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. IEEE Trans. Affect. Comput. (2022). https://doi.org/10.1109/TAFFC.2022.3171091
18. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings CVPR (2021)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings CVPR, pp. 770–778 (2016)
20. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)
21. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings ICML, pp. 8748–8763 (2021)
22. Agarap, A.F.: Deep learning using rectified linear units (ReLU) (2019). arXiv preprint arXiv:1803.08375
23. Sammut, C., Webb, G.I.: Mean squared error. In: Sammut, C., Webb, G.I. (eds.) Encyclopedia of Machine Learning, p. 653. Springer, Boston (2011). https://doi.org/10.1007/978-0-387-30164-8_528
24. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
25. Guntuku, S.C., Lin, W., Carpenter, J., Ng, W.K., Ungar, L.H., Preoţiuc-Pietro, D.: Studying personality through the content of posted and liked images on twitter. In: Proceedings ACM on Web Science Conference, pp. 223–227 (2017)
26. Costa, P.T. Jr.: Revised NEO personality inventory and neo five-factor inventory. Prof. Manual (1992)
27. Zhu, H., Li, L., Zhao, S., Jiang, H.: Evaluating attributed personality traits from scene perception probability. Pattern Recogn. Lett. **116**, 121–126 (2018)