Investigating the Effect of Linguistic Features on Personality and Job Performance Predictions

Hung Le[1(✉)], Sixia Li[1], Candy Olivia Mawalim[1], Hung-Hsuan Huang[2], Chee Wee Leong[3], and Shogo Okada[1]

[1] Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan
{hungle,lisixia,candylim,okada-s}@jaist.ac.jp
[2] The University of Fukuchiyama, Fukuchiyama, Japan
hhhuang@acm.org
[3] Educational Testing Service, Princeton, USA
cleong@ets.org

Abstract. Personality traits are known to have a high correlation with job performance. On the other hand, there is a strong relationship between language and personality. In this paper, we presented a neural network model for inferring personality and hirability. Our model was trained only from linguistic features but achieved good results by incorporating transfer learning and multi-task learning techniques. The model improved the F1 score 5.6% point on the Hiring Recommendation label compared to previous work. The effect of different Automatic Speech Recognition systems on the performance of the models was also shown and discussed. Lastly, our analysis suggested that the model makes better judgments about hirability scores when the personality traits information is not absent.

Keywords: Personality Traits · Job Performance · Social Signal Processing · Natural Language Processing

1 Introduction

The way in which we perceive the world and how the world perceives us is largely influenced by our personality. Psychologists have studied human personalities for many decades, and the Big Five personality model is known as the best working hypothesis [17]. The Big Five model states that human personality differs across five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Emotional Stability (neuroticism). There are many situations where understanding one's personality is beneficial, such as in career coaching or in family conflict resolution. Prior research showed that there is also a strong relationship between a candidate's personality and their job performance [4,13]. Due to these advantages, companies are more and more interested in their candidates' personalities and projected job performance. However, the traditional method to evaluate

A. Coman and S. Vasilache (Eds.): HCII 2023, LNCS 14025, pp. 370–383, 2023.
https://doi.org/10.1007/978-3-031-35915-6_27

personality traits via filling out questionnaire forms is both subjective and time-consuming, so objective methods to automate this process are needed. To this end, machine learning (ML) models are developed for this task in recent years, for their ability to explore human multimodal behaviors.

Previous studies [7,19,21,24] have used both verbal and nonverbal behaviors to predict personality traits. Nonverbal behaviors features were found to be effective in predicting personality and hirability, however, data sets whose nonverbal features are predictive could be suffering from annotations bias (face attractiveness, perceived ages, etc.), and models could unintentionally learn the bias for making predictions [18]. As an attempt to overcome this problem, our work focuses on developing models for only linguistic features extracted from the videos. A neural network (NN) model is proposed to demonstrate that it is possible to simultaneously learn the personalities and job performance of an interviewee from the content of their speech. In addition, not much research [25] was dedicated to studying the relationship between the Big Five personality and the Hiring Recommendation label. In this work, we conducted experiments to show that when giving a model information about the speaker's personality, it could make better predictions about the Hiring Recommendation (hirability) label.

In the field of multimodal learning, Automatic Speech Recognition (ASR) systems are often used to convert speech from audio into text for processing. Different ASR systems may have different performance levels. Word Error Rate (WER), which is defined as the number of incorrectly recognized words divided by the total number of spoken words, is a common metric to evaluate ASR systems. ASRs errors influence NLP and personality trait modeling, but to what extent the influence affects post-processing tasks is not studied well [32]. Therefore, we also conducted experiments with transcriptions obtained from two different ASR systems (namely, Watson and Whisper) and discuss the results.

In summary, the main contributions are:

- A NN model for predicting personality traits and job interview performance is proposed. The model leverages a pre-trained large language model and the weighted linear sum of the losses function to carry out multi-task classification learning. Experimental results showed that the proposed model performed better than previous work when evaluated using the F1 score.
- The effect of two different ASR systems on the performances of different models was analyzed.
- This research is bridging the gap between computers and humans by improving computers' ability to predict human performance in job interview.

2 Related Work

The goal of multimodal machine learning can be defined as "to build models that can process and relate information from multiple modalities" [3]. There are many exciting works on multimodal learning such as works on visual question-answering systems [1,40] or image generation systems [36,37]. One of the fields

that widely uses multimodal learning models is affective computing. In [27], the authors developed a framework to predict job interview performance using facial expressions, language, and prosodic information. [7] developed models to predict personality traits and hiring recommendation scores from monologue videos. [15] collected more than 7000 video job interviews for real positions and developed a hierarchical attention model (HireNet) to predict the hireability of the candidates.

One of the earliest research on computational hirability was conducted by [31]. This work found that it is possible to predict hirability scores from nonverbal features, and the interaction during the interview is more effective for the prediction than psychometric questionnaires data. Following work found that even nonverbal brief excerpts of interactions were still predictive of hirability impressions [29]. For many people, conducting job interviews is a stressful task. The authors of [11] explored the relationship between stress and hirability impressions, and individuals who are perceived as more stressful are more likely to get lower hirability scores. In [30], the authors collected a conversational video resumes dataset and developed a computational framework to predict first impressions, and the analysis showed that there are correlations between personality and hirability. A framework for improving the first impressions of hospitality students was proposed in [26]. Another feature-extraction framework was proposed in [33] to infer personality traits and hiring decisions.

The common features used to train predictive models are linguistic, acoustic, and visual features. Combining multiple modalities does not always mean much better results are obtained, though. In [7], fusing different modalities does not yield better results than models that only have text as the only modality. In [15], the authors stated that "more sophisticated fusion schemes are needed to improve on the monomodal results". One of the possible reasons is that by introducing more modalities, more noise is also introduced, making it more difficult for models to learn the useful signals. In this paper, we focus only on linguistic features to develop the models. This approach was also explored by [9], with the main difference being that our text comes directly from the speeches of interviewees and not from a chat-based interface.

Recent advances in machine learning come largely from the Transformer architecture [39] and its variants. The state of the art of many tasks was raised significantly by models built upon this architecture. In speech recognition, wav2vec2 [2] or Whisper models are approaching human accuracy and robustness. In NLP, large language models such as BERT [10], RoBERTa [23], or GPT-3 [6] do surprisingly well on the text classification task, along with other tasks. That being said, classical machine learning methods such as SVM [8] still have their place as a strong baseline, especially when the classes are clearly separated.

When developing models for large-corpus of conversational videos, it is not practical nor scalable to manually transcript the videos. Instead, an ASR system is usually used to convert speeches to text. In [32], three Japanese ASR systems were compared and their effects on the storytelling skill assessment were evaluated. To the best of our knowledge, no previous work has attempted to evaluate

the effect of different ASR systems on personality traits and interview performance prediction models. Therefore, in this work, we extracted the text from two English ASR systems and analyzed the effect of the ASR error rate on the developed models.

3 Methodology

3.1 Predicting Personality from Linguistic Features

The Big Five traits theory was originally discovered by following the guidelines of the lexical hypothesis, which stated that we use language to encode the difference between people. Therefore, there is a strong relationship between language and personality traits [5]. Automatic personality recognition is one of the important tasks in the Personal Computing research field, as it has many implications in the emerging Human-Centered Artificial Intelligence scenarios.

Feature Representation. Two main approaches used to represent language are the closed-vocabulary approach and the open-vocabulary approach. In the closed-vocabulary approach, words are separated into predefined categories and the correlations between the number of words belonging to each category and personality are studied. The Linguistic Inquiry and Word Count (LIWC) [34] is one of the widely used lexicons in this approach. In the second approach, words and documents are usually converted into vector representations by a language model, and then the vector representations are inputted directly into machine learning models. In the proposed model, we used a large language model (LLM) named RoBERTa as the feature extractor. RoBERTa is an improved version of BERT [10], both of which are pre-trained LLMs based on the Transformer architecture [39]. Unlike the closed-vocabulary approach, where each word has only one concrete meaning, LLMs are capable of taking the word and its surrounding context into account when generating the embedding.

3.2 Dataset and ASR Systems

This work was conducted on the corpus shared by Chen el at. [7]. This corpus contains 1891 monologue videos from 260 interviewees, and each video was annotated with the perceived personality trait and holistic scores. A training set (1519 samples) and a test set (372 samples) were produced under the condition that no interviewee appears in both set. The original scores of the labels were in the 7-point Likert scale, but they are converted into two scores HIGH and LOW using the median scores as the thresholds. Figure 1 shows the training set score distributions of the labels before the conversion, and Table 1 shows the statistics of the labels after the conversion. One key observation from Fig. 1 is that the Hiring Recommendation scores follow the Gaussian distribution, while the Personality Traits scores follow the bimodal distribution.

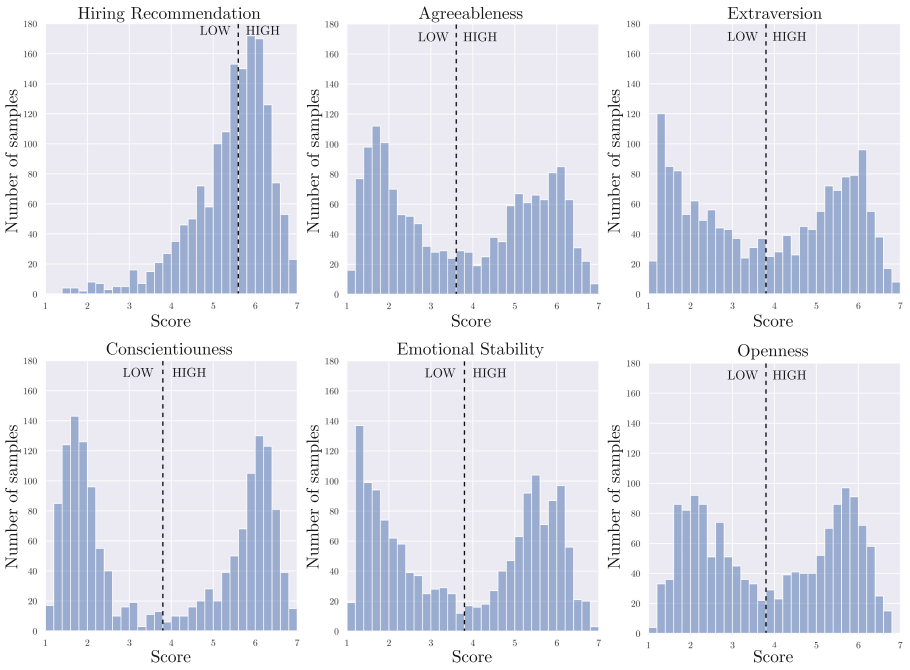

Fig. 1. The distributions of the labels on the training set before converting to HIGH and LOW classes. The dash lines indicate the median scores.

Table 1. The statistics of the labels after converted to HIGH and LOW classes

Label	Training Set		Test Set	
	HIGH	LOW	HIGH	LOW
Hiring Recommendation	773	746	191	181
Agreeableness	780	739	186	186
Extraversion	774	745	189	183
Conscientiousness	761	758	186	186
Emotional Stability	781	738	180	192
Openness	788	731	188	184

To convert the candidates' speech to text, two ASR systems were chosen: IBM Watson[1] (commercially available, transcription files were provided by the original authors of [7]), and OpenAI's Whisper system [35]. The WERs of the systems when calculated based on 22 manually transcribed random samples from the data set are 32.73% and 5.28%, and the average lengths of the transcriptions are 275 and 276 words, respectively.

[1] https://www.ibm.com/cloud/watson-text-to-speech.

3.3 Proposed Model: The Neural Network

In this paper, a neural network (NN) model was proposed as an alternative approach to the baseline model (the Support Vector Machine, to be introduced in Subsect. 3.4). Figure 2 shows the architecture of our model. Two main advantages of the neural network model compares to the baseline model are: more information can be encoded to the text embedding vector by the pre-trained language model, and only one model is trained for all six labels.

Since each candidate was given two minutes to answer a question, the answers came in form of a paragraph. We feed the paragraphs to RoBERTa [23] to obtain the paragraph embeddings. In this paper, we used the "roberta-base"[2] version, with the maximum input length set to 512 tokens. The output of this step is an embedding vector that has 768 dimensions. This embedding vector is then concatenated with the unique z-normalized numbers indicating the Speaker ID and the Question ID, resulting in a vector that has 770 dimensions (similar to the models proposed by [28]). The original (unnormalized) Speaker ID and Question ID are two unique integers indicating which speaker (interviewee) is answering which question. Since there are a total of 260 speakers and a maximum of 8 questions, the original Speaker ID ranges between 1 and 260, while the original Question ID ranges between 1 and 8. The Speaker ID and Question ID are inputted to the NN to provide the model with additional contextual information about the paragraph.

The 770 dimensions vector is then passed to four blocks of layers. Each of the first 3 blocks contains a Fully Connected (FC) layer, a LeakyReLU nonlinear activation layer, and a Dropout layer with a dropout probability of 0.5. The last block contains only a Fully Connected layer (FC_4). The first three FC layers are initialized using Kaiming initialization [14], while the FC_4 layer is initialized using Xavier initialization [12]. The final output is a vector that has 6 dimensions, corresponding to the six labels (the Hiring Recommendation and the Big Five personality traits).

The Loss Function. Since we formulated the problem as a multi-label classification problem, the Binary Cross-Entropy (BCE) loss is the natural choice for the loss function. However, the experiments showed that when BCE loss is naively applied, the model performs well on the Big Five labels, while performing poorly on the Hiring Recommendation label. The results suggest that the Hiring Recommendation label is more difficult to classify compared to other labels. Therefore, we modified the BCE loss to the weighted linear sum of the losses, which takes the formula:

$$\mathcal{L}_{total} = \sum_i w_i \mathcal{L}_i \tag{1}$$

where i is the label, $\mathcal{L}_i$ is the BCE loss with respect to label i, and w_i is the weighted parameter for $\mathcal{L}_i$. This loss function was previously used in multi-task

[2] https://huggingface.co/roberta-base.

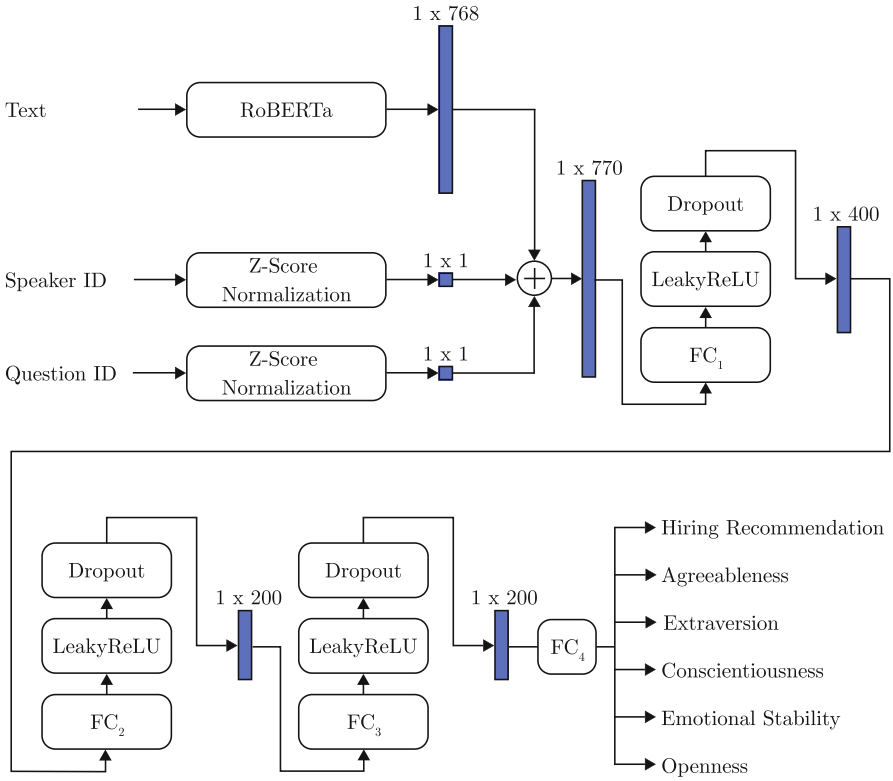

Fig. 2. The architecture of the neural network model ("FC" denotes the "Fully Connected" layer)

learning problems [20,22,38]. In our case, the concrete formula is:

$$\mathcal{L}_{total} = w_{hr}\mathcal{L}_{hr} + w_{ag}\mathcal{L}_{ag} + w_{ex}\mathcal{L}_{ex} + w_{co}\mathcal{L}_{co} + w_{em}\mathcal{L}_{em} + w_{op}\mathcal{L}_{op} \quad (2)$$

where the subscripts (hr, ag, ex, co, em, op) stand for the labels (hiring recommendation, agreeableness, extraversion, conscientiousness, emotional stability, openness), respectively.

Hyperparameters. Grid search was used for selecting the NN hyperparameters. To perform grid-search, roughly 20 percent of the original training set (298 samples) was separated to create the validation set. The separation was also performed under the condition that no speaker appears in both sets. Table 2 shows the best hyperparameters for the NN model.

Table 2. Training hyperparameters

Hyperparameter	Value
Batch Size	32
Optimizer	AdamW
β_1	0.9
β_2	0.99
ϵ	$1e-6$
Weight Decay	$1e-2$
AMSGrad	True
Training epoch	500
Max Learning Rate	$1e-2$
Learning Rate Scheduler	Cosine Annealing with Hard Restarts and Warm up
Number of restart cycles	2
Total training steps	24000
Warm-up steps	2400

3.4 Baseline Model: The Support Vector Machine

The Support Vector Machine (SVM) was chosen as the baseline model since this approach produced the best classifiers for this data set in previous work [7]. The general pipeline showed in [7] was followed: features are extracted from text using the Bag-of-Words model, then feed into the SVM. The Radius Basic Function (RBF) kernel was used in this study and grid-search was used for selecting the parameters C and γ from the following range [16]:

$$C \in \{2^{-5}, 2^{-3}, 2^{-1}, 1, 2^1, 2^3, 2^5, 2^7, 2^9\}$$

$$\gamma \in \{2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^1, 2^3\}$$

The best parameters were chosen using 5-fold cross-validation on the training set, where the folds were separated under the same condition that the test set was separated.

4 Results

For the evaluation metric, the macro F1 measurement is reported (we found that the Precision and Recall scores are mostly equal to F1). Table 3 shows the experiment results of our methods. For the NN model, the table shows the average results of 5 runs with different random seeds.

Table 3. F1 scores on the test set

Label	Model			
	SVM		Neural Network	
	ASR System		**ASR System**	
	Watson (obtained from [7])	Whisper (ours)	Watson (ours)	Whisper (ours)
Hiring Recommendation	0.66	0.69	**0.716**	0.714
Agreeableness	0.84	0.85	0.838	**0.862**
Extraversion	0.78	0.80	**0.812**	0.802
Conscientiousness	0.86	0.86	0.85	**0.868**
Emotional Stability	0.83	0.84	0.85	**0.88**
Openness	0.81	0.83	0.828	**0.844**

5 Analysis and Discussion

Performance of the Proposed Model Compared to the Baseline Model. Table 3 shows that the performance of the proposed model is higher than the baseline models across all labels. The best results mostly come from the NN model trained on transcriptions from the ASR system with the lowest WER (Whisper). The highest F1 score for the Hiring Recommendation label is 0.716, which is a 0.056-point increment compared to previous work. For the Big Five personality trait labels, the gains range between 0.01 to 0.05 points.

Interpretation for the Improved Performance of the NN Model. When human annotators annotated the original videos, they did not watch the videos and annotated each of the labels separately. Instead, they watched a video once, and then annotated all the labels. Differing from the SVM models, the multi-label NN reassembled this process closely. In the SVM baseline, each model is separately trained with respect to each of the labels, so the Hiring Recommendation prediction model does not have access to the Personality Trait labels. On the other hand, the proposed NN model updated its weights from the feedback of all the labels at the same time, so the NN model can learn some relationships between the labels. To evaluate the effect of the Personality Trait labels on the model's ability to predict the Hiring Recommendation score, we conducted experiments with some changes to the weights of the loss function. In particular, we set the weights of the Personality Trait labels in Eq. 2 to zeros. We retrained the NN model on the Whisper's transcriptions and found that the 5-run average F1 score of the Hiring Recommendation label decreased to 0.696. This is similar to the results of the baseline model, where the Personality Trait labels also were not taken into account.

The Effect of ASR Systems on Models' Performance. With respect to different ASR systems, the results show that the SVM models benefited from the higher quality transcriptions, while it is not clear that the NN model received the same benefits. In the case of Hiring Recommendation and Extraversion labels, the NN model performed slightly worse when trained on higher-quality transcriptions. It is possible that the negative gains simply come from the randomization nature of NN models.

Sum of the BCE Losses. In Sect. 3.3, we mentioned that the weighted linear sum of the BCE losses function helped the model learn all labels efficiently. In this section, more details to support the claim are provided. By conducting parameter searches, our experiments show that when w_{hr} in Eq. 2 is 5 and w of each of the Big Five labels loss is 1, the model is able to learn all labels simultaneously. On the other hand, when all the weights in Eq. 2 are set to 1 (called the linear sum of the BCE losses function), the Hiring Recommendation label cannot be learned. Figure 3 shows the accuracy of the Hiring Recommendation label (the model was trained on Whisper's transcriptions) on the validation set when the weights are set in the two cases. The figure does not show the accuracy of the other labels since those are almost identical.

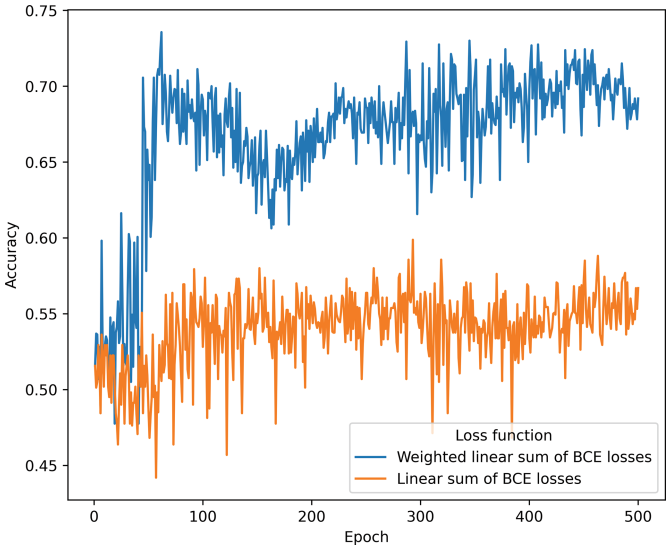

Fig. 3. The accuracy of the Hiring Recommendation label on the validation set when different weights are used for the loss function. The validation accuracy of the Big Five labels is omitted to simplify the figure.

6 Conclusion

In this work, we proposed a NN model for predicting personality traits and hiring recommendation scores. The experiment results showed that our NN architecture performs better than the baseline model. While the Big Five labels can be predicted quite accurately, predicting whether a candidate should be invited to an onsite interview is a much more challenging task. Our analysis showed that it is better to give the model can learn some relations between personality traits and the hiring recommendation labels. The effect of distinct ASR systems on the models' performances was also evaluated. We also found that the quality of the transcriptions only has little effect on the models' performance.

There are still some limitations to our approach. First, other modalities besides text were not considered. Secondly, our work is based on the assumption that the way people use their language in front of the camera is similar to real life. This is not always the case. In future work, we plan to incorporate other modalities such as visual and acoustic modalities into the NN model. Furthermore, in the context of HCI, more research is needed to understand how different types of people use their language differently in front of cameras.

Acknowledgements. This work was also partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (No. 22K21304, No. 22H04860 and 22H00536), JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6) and JST Moonshot R&D program (JPMJMS2237-3). Hung is supported by the Japanese Government (MEXT) Scholarship.

References

1. Antol, S., et al.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015)
2. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations (2020). https://doi.org/10.48550/ARXIV.2006.11477
3. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 423–443 (2019)
4. Barrick, M.R., Mount, M.K.: The big five personality dimensions and job performance: A meta-analysis. Pers. Psychol. **44**(1), 1–26 (1991). https://doi.org/10.1111/j.1744-6570.1991.tb00688.x
5. Boyd, R.L., Pennebaker, J.W.: Language-based personality: a new approach to personality in a digital world. Current Opinion in Behavioral Sciences **18**, 63–68 (2017). https://doi.org/10.1016/j.cobeha.2017.07.017, big data in the behavioural sciences
6. Brown, T.B., et al.: Language models are few-shot learners (2020). https://doi.org/10.48550/ARXIV.2005.14165
7. Chen, L., Zhao, R., Leong, C.W., Lehman, B., Feng, G., Hoque, M.E.: Automated video interview judgment on a large-sized corpus collected online. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 504–509 (2017). https://doi.org/10.1109/ACII.2017.8273646

8. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (sep 1995). https://doi.org/10.1023/A:1022627411411
9. Dai, Y., Jayaratne, M., Jayatilleke, B.: Explainable personality prediction using answers to open-ended interview questions. Front. Psychol. **13** (2022). https://doi.org/10.3389/fpsyg.2022.865841
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). https://doi.org/10.48550/ARXIV.1810.04805
11. Finnerty, A.N., Muralidhar, S., Nguyen, L.S., Pianesi, F., Gatica-Perez, D.: Stressful first impressions in job interviews. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 325–332. ICMI '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2993148.2993198
12. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 249–256. PMLR, Chia Laguna Resort, Sardinia, Italy (13–15 May 2010)
13. Goodstein, L.D., Lanyon, R.I.: Applications of personality assessment to the workplace: a review. J. Bus. Psychol. **13**, 291–322 (1999). https://doi.org/10.1023/A:1022941331649
14. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034 (2015). https://doi.org/10.1109/ICCV.2015.123
15. Hemamou, L., Felhi, G., Vandenbussche, V., Martin, J.C., Clavel, C.: Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 573–581 (07 2019). https://doi.org/10.1609/aaai.v33i01.3301573
16. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University (2003). http://www.csie.ntu.edu.tw/ cjlin/papers.html
17. John, O.P., Srivastava, S.: The big five trait taxonomy: History, measurement, and theoretical perspectives (1999)
18. Junior, J.C.S.J., Lapedriza, A., Palmero, C., Baró, X., Escalera, S.: Person perception biases exposed: Revisiting the first impressions dataset. In: 2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW), pp. 13–21 (2021). https://doi.org/10.1109/WACVW52041.2021.00006
19. Katada, S., Okada, S.: Biosignal-based user-independent recognition of emotion and personality with importance weighting. Multimedia Tools Appl. **81**(21), 30219–30241 (sep 2022). https://doi.org/10.1007/s11042-022-12711-8
20. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics (2017). https://doi.org/10.48550/ARXIV.1705.07115
21. Kwon, S., Choeh, J.Y., Lee, J.W.: User-personality classification based on the non-verbal cues from spoken conversations. Int. J. Comput. Intell. Syst. **6**, 739–749 (05 2013). https://doi.org/10.1080/18756891.2013.804143
22. Liao, Y., Kodagoda, S., Wang, Y., Shi, L., Liu, Y.: Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 2318–2325. IEEE Press (2016). https://doi.org/10.1109/ICRA.2016.7487381

23. Liu, Y., et al.: Roberta: A robustly optimized bert pretraining approach (2019). https://doi.org/10.48550/ARXIV.1907.11692
24. Mawalim, C.O., Okada, S., Nakano, Y.I., Unoki, M.: Multimodal bigfive personality trait analysis using communication skill indices and multiple discussion types dataset. In: Meiselwitz, G. (ed.) Social Computing and Social Media. Design, Human Behavior and Analytics - 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, IN: Proceedings, Part I. Lecture Notes in Computer Science, vol. 11578, pp. 370–383. Springer (2019). https://doi.org/10.1007/978-3-030-21902-4_27
25. Mujtaba, D.F., Mahapatra, N.R.: Multi-task deep neural networks for multimodal personality trait prediction. In: 2021 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 85–91 (2021). https://doi.org/10.1109/CSCI54926.2021.00089
26. Muralidhar, S., Nguyen, L.S., Frauendorfer, D., Odobez, J.M., Schmid Mast, M., Gatica-Perez, D.: Training on the job: Behavioral analysis of job interviews in hospitality. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 84–91. ICMI '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2993148.2993191
27. Naim, I., Tanveer, M., Gildea, D., Hoque, E.: Automated prediction and analysis of job interview performance: The role of what you say and how you say it (05 2015). https://doi.org/10.1109/FG.2015.7163127
28. Nakano, Y.I., Hirose, E., Sakato, T., Okada, S., Martin, J.C.: Detecting change talk in motivational interviewing using verbal and facial information. In: Proceedings of the 2022 International Conference on Multimodal Interaction, pp. 5–14. ICMI '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3536221.3556607
29. Nguyen, L., Gatica-Perez, D.: I would hire you in a minute: Thin slices of nonverbal behavior in job interviews, pp. 51–58 (11 2015). https://doi.org/10.1145/2818346.2820760
30. Nguyen, L., Gatica-Perez, D.: Hirability in the wild: analysis of online conversational video resumes. IEEE Trans. Multimed. **18**, 1422–1437 (07 2016). https://doi.org/10.1109/TMM.2016.2557058
31. Nguyen, L.S., Frauendorfer, D., Mast, M.S., Gatica-Perez, D.: Hire me: computational inference of hirability in employment interviews based on nonverbal behavior. IEEE Trans. Multimedia **16**(4), 1018–1031 (2014). https://doi.org/10.1109/TMM.2014.2307169
32. Okada, S., Komatani, K.: Investigating effectiveness of linguistic features based on speech recognition for storytelling skill assessment. In: Mouhoub, M., Sadaoui, S., Ait Mohamed, O., Ali, M. (eds.) Recent Trends and Future Technology in Applied Intelligence, pp. 148–157. Springer International Publishing, Cham (2018)
33. Okada, S., Nguyen, L., Aran, O., Gatica-Perez, D.: Modeling dyadic and group impressions with intermodal and interperson features. ACM Trans. Multimed. Comput., Commun. Appl. **15**, 1–30 (01 2019). https://doi.org/10.1145/3265754
34. Pennebaker, J., Boyd, R., Jordan, K., Blackburn, K.: The development and psychometric properties of LIWC2015. University of Texas at Austin (2015). https://doi.org/10.15781/T29G6Z
35. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022). https://doi.org/10.48550/ARXIV.2212.04356

36. Ramesh, A., et al.: Zero-shot text-to-image generation (2021) arxiv:2102.12092
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
38. Uhrig, J., Cordts, M., Franke, U., Brox, T.: Pixel-level encoding and depth layering for instance-level semantic labeling. In: German Conference on Pattern Recognition (2016)
39. Vaswani, A., et al: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
40. Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., Van Den Hengel, A.: Visual question answering: a survey of methods and datasets. Comput. Vis. Image Underst. **163**, 21–40 (2017)