Interspeech 2018
2-6 September 2018, Hyderabad

Lexical And Acoustic Deep Learning Model For Personality Recognition

Guozhen An[1,2], *Rivka Levitan*[1,3]

[1]Department of Computer Science, CUNY Graduate Center, USA
[2]Department of Mathematics and Computer Science, York College (CUNY), USA
[3]Department of Computer and Information Science, Brooklyn College (CUNY), USA

gan@gradcenter.cuny.edu, rlevitan@brooklyn.cuny.edu

Abstract

Deep learning has been very successful on labeling tasks such as image classification and neural network modeling, but there has not yet been much work on using deep learning for automatic personality recognition. In this study, we propose two deep learning structures for the task of personality recognition using acoustic-prosodic, psycholinguistic, and lexical features, and present empirical results of several experimental configurations, including a cross-corpus condition to evaluate robustness. Our best models match or outperform state-of-the-art on the well-known myPersonality corpus, and also set a new state-of-the-art performance on the more difficult CXD corpus.

Index Terms: Personality recognition, Deception detection, DNN, LSTM, Word Embedding

1. Introduction

Automatic personality recognition is useful for many computational applications, including recommendation systems, dating websites, and adaptive dialogue systems, as a feature that can both inform personalization and help predict useful information such as job performance or academic outcomes.

Personality refers to individual differences in characteristic patterns of thinking, feeling, and behaving [1]. A commonly used model of personality is the NEO-FFI five factor model of personality traits, also known as the Big Five: Openness to Experience (having wide interests, imaginative, insightful), Conscientiousness (organized, thorough, a planner), Extroversion (talkative, energetic, assertive), Agreeableness (sympathetic, kind, affectionate), and Neuroticism (tense, moody, anxious) [2]. These traits were originally identified by several researchers working independently [3] and the model has been employed to characterize personality in multiple cultures [4].

Most previous research on personality detection has used personality scores assigned by annotators based solely on the text or audio clip. While such annotations can be useful for studying how personality is *perceived*, they have been shown to correlate only weakly with scores derived from personality tests completed by the subject themselves ("self-reported" labels), and have moderate to weak internal consistency (as measured by Cronbach's alpha) [5]. Predicting self-reported NEO-FFI scores, as we do here, is a much more difficult task (since the stranger ratings are based only on the speech or text samples, which necessarily contain all the information needed for the prediction).

The main contribution of this paper is a multi-modal deep learning model for personality prediction using acoustic-prosodic features and word embeddings. We experimented with a standard multi-layer perceptron (MLP) as well as a model that uses an LSTM (long short term memory) layer to encode each instance's word vectors for the final prediction. For the standard MLP approach, we further experiment with methods for combining feature sets of different sizes and modalities.

For comparison with previous work, we report the performance of each model on the myPersonality corpus [6], which has only text data, as well as the CXD corpus [7, 8], which contains audio and text. Additionally, we explore the ability of our model to represent personality in general, rather than on a specific corpus, by testing the performance of each trained model on its opposite corpus.

The remainder of this paper is structured as follows. In Section 2, we review previous work. Description of the dataset can be found in Section 3. In Section 4, we described feature set and model detail. Section 5 present the experimental setup and results from various models. Finally, We conclude and discuss future research directions in Section 6.

2. Related Work

There have been numerous successful approaches to the automatic personality recognition in the literature, using various combinations of acoustic-prosodic, lexical, and psycholinguistic features [9, 10, 11, 12, 13, 14]. [10] used prosodic features to detect personality from ten second audio clips labeled for personality by human judges based on the audio clips alone. [9] used lexical features to predict personality traits in student essays. Features based on Linguistic Inquiry and Word Count (LIWC) [11] psycholinguistic categories have been shown to correlate with Big Five personality traits, both in writing samples [15] and in spoken dialogue [12]. [13] used acoustic-prosodic, lexical, and psycholinguistic features to predict self-reported personality labels in deceptive interviews. The results in [14] predicted self and observer personality scores from essay and conversational data, using LIWC, psycholinguistic, and prosodic feature sets. Their results indicate that observer reports are easier to predict: their models predicted observer labels with good accuracy but did not outperform the baseline for self-reported labels.

Most approaches to personality prediction have used traditional machine learning algorithms such as Support Vector Machine and Naive Bayes. Recently, several studies have applied deep learning, which has achieved groundbreaking results in many areas, to the task of personality prediction.

[16] used convolutional filters to aggregate word vectors into sentence vectors, and used those features together with the features used by [14] in a multilayer perception (MLP) to predict self-reported personality labels on a stream-of-consciousness essay dataset [15], achieving an average of approximately 58% accuracy on the Big Five personality traits, using different configurations to achieve the best results for each trait.

10.21437/Interspeech.2018-2263

[17] used a multilayer perception on the myPersonality dataset [6] with GloVe word embeddings [18]. They achieved approximately 71% average accuracy on all traits, performing as high as 79% on Openness and Neuroticism.

Finally, [19] built a classifier for personality recognition using a convolutional neural network (CNN) with bilingual word embeddings, also on the myPersonality data, with approximately 66% average accuracy.

Our work is differentiated from previous work by the inclusion of acoustic-prosodic features, and the proposal and evaluation of three architectures for combining features from different modalities. Furthermore, we predict personality labels from much more challenging data: unlike the myPersonality and essay datasets, the CXD dataset, described more fully in Section 3, is spoken, task-oriented, and deceptive.

Omitting the acoustic-prosodic features, we evaluate our models on the myPersonality dataset for the sake of direct comparison with previous work. We cannot directly compare our accuracy, with respect to a binary classification, to previous work on the CXD dataset, which predicted high/medium/low labels [13, 20], but we trained an SVM for binary predictions with the same set of features for comparison with that work.

3. Data

Two labeled personality datasets were used for our study. The first dataset was collected by myPersonality project [6]. The Facebook dataset contains 9917 status updates in raw text from 250 Facebook users. Gold standard Big Five personality labels were obtained for each user using an 100-item long version of the IPIP personality questionnaire. Both scores and classes were included in the dataset, and classes have derived from scores with a median split.

Table 1: *User-level personality distribution of myPersonality dataset*

Value	O	C	E	A	N
Yes	176	130	96	134	99
No	74	120	154	116	151

Table 2: *Status-level personality distribution of myPersonality dataset*

Value	O	C	E	A	N
Yes	7370	4556	4210	5268	3717
No	2547	5361	5707	4649	6200

The second dataset was collected by the Deception project at Columbia University, and the collection and design of the corpus analyzed here is described in more detail by [7, 21, 22]. It contains within-subject deceptive and non-deceptive English speech, collected using a fake resume paradigm, from native speakers of Standard American English (SAE) and Mandarin Chinese (MC). There are approximately 125 hours of speech in the corpus from 173 subject pairs and 346 individual speakers. Big Five personality scores were obtained for each speaker using the NEO-FFI personality inventory [2]. Classes were derived by splitting scores at the median.

The unit of segmentation used here is the turn. Turn boundaries were extracted in the following manner: the manual orthographic transcription was force-aligned with the audio, and the speech was segmented if there was a silence of more than 0.5 seconds. In total, there are 29175 turn-level instances. The average duration of each instance is 9.03s, though there are quite a few outliers. During training and testing, all turns from a single speaker were contained within a single fold.

Table 3: *Status-level distribution of CXD dataset*

Value	O	C	E	A	N
Yes	16289	14593	15504	15653	14783
No	12886	14582	13671	13522	14392

4. Methodology

4.1. Features

For our experiments, we use the feature sets described in [13]: acoustic-prosodic low-level descriptor features (**LLD**); word category features from **LIWC** (Linguistic Inquiry and Word Count) [11]; and word scores for pleasantness, activation and imagery from the Dictionary of Affect in Language (**DAL**) [23]. We also add two new feature sets based on word embeddings.

Low-Level Descriptor (LLD). The Low-Level Descriptor (LLD) feature set contains approximately 384 acoustic-prosodic features as described in the Interspeech 2009 COMPARE Challenge [24]. These are with extracted using the baseline 2009 Challenge configuration. The Low-Level Descriptor features include pitch (fundamental frequency), intensity (energy), spectral, cepstral (MFCC), duration, voice quality (jitter, shimmer, and harmonics-to-noise ratio), spectral harmonicity, and psychoacoustic spectral sharpness.

Linguistic Inquiry and Word Count (LIWC). We used Linguistic Inquiry and Word Count (LIWC) [11] to extract the lexical features. LIWC is a text analysis program that calculates the degree to which people use different categories of words, and can determine the degree any text uses positive or negative emotions, self-references, causal words, and 70 other language dimensions. We extracted a total of 130 LIWC features based on the 64 LIWC categories: 64 features based upon the ratio of words appearing in each LIWC categories over total word count; 64 features based on the ratio of words appearing in each LIWC categories over the total words appearing in any LIWC category; the total number of words appearing in any LIWC category; and the total word count.

Dictionary of Affect in Language (DAL). We used Whissell's Dictionary of Affect in Language (DAL) [23] to extract additional features. The DAL is a lexical analysis tool which is used for analyzing emotive content of speech especially for *pleasantness*, *activation* and *imagery*. It lists approximately 4500 English words, each with ratings for these three categories in the DAL. These were obtained from multiple human judges. We extract 19 features derived from the DAL scores for each word in each subject's baseline interview transcript. From all words' pleasantness, activation and imagery scores, we calculated the mean, minimum, maximum, median, standard deviation, and variance. We also added the number of words in the transcript that appear in the DAL.

Word embeddings (WE). We use the Gensim library [25] to extract two sets of word vector features using Google's pre-trained skip-gram vectors [26] and Stanford's pre-trained GloVe vectors [18]. In order to calculate the vector representation of a turn, we extract a 300-dimensional word vector for each word in the segment segment, and then average them to get a 300-dimensional vector representing the entire turn segment.

The feature sets used here represent information from both the acoustic and lexical signal, as well has the higher-level psycholinguistic information represented by the LIWC and DAL features. They also vary widely in size, from 19 features (DAL) to 384 (LLD). We therefore experiment with several methods for combining feature sets from different modalities, described in more detail in Section 4.2.

4.2. Multilayer perceptron

Our first model is a multilayer perceptron (MLP) [27], a simple feed-forward network using the sigmoid activation function. We use two different approaches to combine the feature sets.

First, we try an early-fusion approach, concatenating all feature sets into a single input feature vector (Figure 1). The network has five fully-connected layers in a bottleneck configuration: (2048, 1024, 512, 1024, 2048) neurons per layer.

Second, we try late fusion, feeding each feature set separately to an individual MLP, and concatenate the output layers to predict each personality traits (Figure 2). Networks with three fully-connected layers of size (256, 128, 256) were used for the DAL, Google WV and GloVe WV feature sets, and size (512, 256, 512) for LIWC and LLD. After the individual MLPs were trained, the last fully-connected layers from each one were concatenated together and fed forward to an output layer with five neurons, one for each trait. This approach balances the influence of each of the feature sets so that a large but possibly less informative feature set does not overwhelm the other features.

Figure 1: *Diagram of first MLP model. LLD was used only for deception dataset*

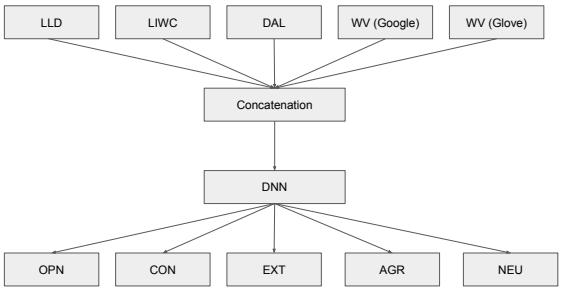
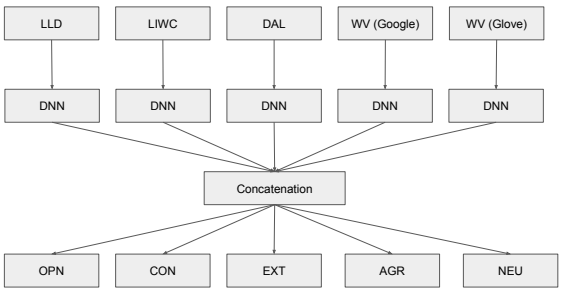


Figure 2: *Diagram of second MLP model. LLD was used only for deception dataset*

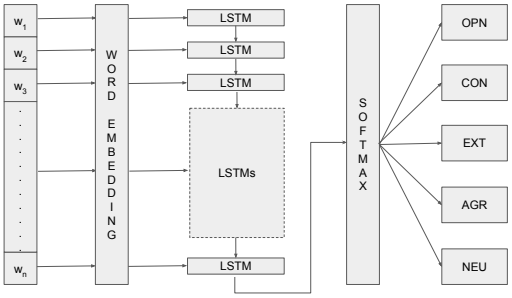


In addition to the sigmoid activation function, we also experimented with ReLU and tanh, but did not see an increase in performance. Tuning the learning rate improved the performance, and the final model used $\alpha = 0.001$ with 100 epochs. For the loss function, we used mean squared error.

4.3. Word Embedding and LSTM

In the models described above, we represented an instance's lexical content by averaging together its word vectors. This approach is quite common but naive. We additionally experiment with feeding an instance's word embeddings into an LSTM (Long Short Term Memory) layer, well known for capturing sequential information [28, 29], to learn an instance-level representation.

Figure 3: *Diagram of WE-LSTM model.*



We also updated off-the-shelf word embeddings to better represent our data. We initialized a 300-dimensional word embedding layer with the GloVe off-the-shelf embeddings. We then trained the new model on our data. Since our corpora are relatively small, this took advantage of the enormous corpora that were used to train the off-the-shelf embeddings, and adapted them to our data.

After training the word embedding layer, we feed 300-dimensional word embeddings one at a time to the LSTM layer to get instance-level representations. We set the maximum word length of each corpus to 60, and zero padding is used if the sentence length is less than 60 words. The LSTM layer's output, which represents the instance's lexical content, is a 256-dimensional vector.

A softmax function is then applied to the instance representation, outputting a probability estimation of the binary classification of each personality trait. We set the learning rate to 0.001, and we use mean squared error loss function.

5. Result

Key results are presented in Tables 4 and 5. We tested our models on both the myPersonality and CXD corpora, as described in Section 3. We compare our results to traditional machine learning (ML) approaches as well as published state-of-the-art on the myPersonality dataset. Traditional ML experiments were done using 10-fold cross-validation, and the deep learning experiments used a 90%/10% train/test split.

We tested various traditional ML algorithms. Like [17], we found that LDA (linear discriminant analysis) performed best on the myPersonality data, although our average accuracy was 61% compared to their published 63%. The discrepancy can be explained by the fact that we omit a data pre-processing step that they applied; furthermore, their published average accuracy aggregates the best results for each personality trait across different experimental configurations. A decision tree had the best

Table 4: *Key results: myPersonality*

Model	O	C	E	A	N	Avg
LDA	.73	.57	.57	.56	.61	.61
s-o-a[1]	**.79**	.59	**.79**	.56	**.79**	**.71**
MLP-1	.76	.60	.61	**.61**	.65	.65
MLP-2	.76	.62	.61	.60	.65	.65
LSTM	.76	.62	.63	.60	.65	.65
MLP-LSTM	.77	**.63**	.64	**.61**	.68	.67

[1] state-of-the-art: [17].

Table 5: *Key results: CXD*

Model	O	C	E	A	N	Avg
Decision tree	.46	.47	.50	.56	.52	.50
MLP-1	.58	**.61**	.52	**.64**	.55	.58
MLP-2	**.60**	**.61**	.52	**.64**	.57	.59
MLP-1+2	**.60**	**.61**	**.59**	**.64**	**.61**	**.61**
LSTM	.59	.59	.51	.60	.53	.56
MLP-LSTM	**.60**	.58	.51	**.64**	.54	.57

performance for the CXD data, with average accuracy of only 50% – essentially random. This confirms that the CXD data presents a significantly more difficult task.

In addition to the two MLP structures and LSTM structure described in Sections 4.2 and 4.3, we tested a model that learned a linear combination of the predictions made by both the MLP and LSTM to produce a final fused prediction. This model performed best for the myPersonality data, with an average accuracy of 67%. This outperforms other recently published results on the same data ([19], 65%). [17] reported 71% accuracy. As with the LDA model, we believe the discrepancy may be explained by the preprocessing and resampling steps not implemented here. Another potential reason for underperforming than their model on openness, extraversion and neuroticism traits is that myPersonality dataset was highly unbalanced on these three trait. Therefore, it is easy to perform better on those three traits by predicting majority class. Instead, our model performs better on the conscientiousness and agreeableness traits, for which the majority baseline is lower than it is for the other traits.

For the CXD corpus, the best model was the combination of two MLP models, with 61% average accuracy, though the early fusion MLP model and late fusion MLP model had individual accuracies of 58% and 59%, respectively. This gives a 11% absolute improvement over the best traditional ML model.

In order to assess whether these models are generalizable across domains for personality recognition, we experiment with training on one dataset and testing on another. For both datasets, the MLPs performed best: 60% average accuracy for myPersonality and 53% for CXD. Both these scores are significantly worse than the best scores of the within-corpus condition, but these accuracies still exceed (CXD) or match (myPersonality) the performances of traditional ML models in the within-corpus condition. We conclude that these models do not generalize well, but can still capture useful information across corpora.

Literature on other tasks suggests that multi-task learning (MTL) – using a single model to predict multiple labels – increases performance. MTL is implemented naturally in neural networks by including several nodes in the output layer, which was done throughout this study. However, when we isolated the effect of MTL by training individual MLPs for each trait, we did not see a significant drop in performance.

6. Conclusion and Future Work

In this paper, we present a deep learning approach to personality recognition that outperforms traditional ML models and recent deep learning models. We compared two network structures, MLP and WE-LSTM, and showed that a network based on the combination of the two performs best on the myPersonality corpus. The MLP, in contrast, generalizes better across corpora, and also performs better on the CXD corpus, which contains many out-of-vocabulary words from the speakers for whom English is a second language. This points to the promise of acoustic-prosodic features, which are more robust with respect to language, and have not previously been used with a deep neural network to predict personality. Finally, we show that early- and late-fusion MLP models achieve comparable performance, though the late-fusion MLP performs better for Openness and Neuroticism in CXD.

In future work we will explore the fusion of these findings to other dataset and other related problems, such as deception detection. We also see the potential of extending our framework to speech signal instead of text for personality recognition.

7. Acknowledgments

This work was partially funded by AFOSR FA9550-11-1-0120.

8. References

[1] A. E. Kazdin, "Encyclopedia of psychology," 2000.

[2] P. T. Costa and R. R. MacCrae, *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*. Psychological Assessment Resources, 1992.

[3] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual review of psychology*, vol. 41, no. 1, pp. 417–440, 1990.

[4] R. R. McCrae, "Trait psychology and culture: Exploring intercultural comparisons," *Journal of personality*, vol. 69, no. 6, pp. 819–846, 2001.

[5] P. Borkenau and A. Liebler, "Trait inferences: Sources of validity at zero acquaintance." *Journal of Personality and Social Psychology*, vol. 62, no. 4, p. 645, 1992.

[6] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines." *American Psychologist*, vol. 70, no. 6, p. 543, 2015.

[7] S. I. Levitan, M. Levine, J. Hirschberg, N. Cestero, G. An, and A. Rosenberg, "Individual differences in deception and deception detection," 2015.

[8] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Speech Prosody*, 2016.

[9] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type," in *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, 2005.

[10] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proceedings of the 2nd international workshop on Social signal processing*. ACM, 2010, pp. 17–20.

[11] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, p. 2001, 2001.

[12] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, "Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life." *Journal of personality and social psychology*, vol. 90, no. 5, p. 862, 2006.

[13] G. An, S. I. Levitan, R. Levitan, A. Rosenberg, M. Levine, and J. Hirschberg, "Automatically classifying self-rated personality scores from speech," *Interspeech 2016*, pp. 1412–1416, 2016.

[14] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, pp. 457–500, 2007.

[15] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

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

[17] T. Tandera, D. Suhartono, R. Wongso, Y. L. Prasetio *et al.*, "Personality prediction system from facebook users," *Procedia Computer Science*, vol. 116, pp. 604–611, 2017.

[18] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[19] F. B. Siddique and P. Fung, "Bilingual word embeddings for cross-lingual personality recognition using convolutional neural nets," *Learning*, vol. 21, p. 22, 2017.

[20] G. An and R. Levitan, "Comparing approaches for mitigating intergroup variability in personality recognition," *arXiv preprint arXiv:1802.01405*, 2018.

[21] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, "Cross-cultural production and detection of deception from speech," in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, 2015, pp. 1–8.

[22] S. I. Levitan, Y. Levitan, G. An, M. Levine, R. Levitan, A. Rosenberg, and J. Hirschberg, "Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection," in *Proceedings of NAACL-HLT*, 2016, pp. 40–44.

[23] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec, "A dictionary of affect in language: Iv. reliability, validity, and applications," *Perceptual and Motor Skills*, vol. 62, no. 3, pp. 875–888, 1986.

[24] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[25] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[27] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14, pp. 2627–2636, 1998.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] Y. Xian and Y. Tian, "Self-guiding multimodal lstm-when we do not have a perfect training dataset for image captioning," *arXiv preprint arXiv:1709.05038*, 2017.