Deep Personality Recognition For Deception Detection

Guozhen An[1,2], *Sarah Ita Levitan*[4], *Julia Hirschberg*[4], *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
[4]Department of Computer Science, Columbia University, USA

gan@gradcenter.cuny.edu, sarahita@cs.columbia.edu
julia@cs.columbia.edu, rlevitan@brooklyn.cuny.edu

Abstract

Researchers in both psychology and computer science have suggested that modeling individual differences may improve the performance of automatic deception detection systems. In this study, we fuse a personality classification task with a deception classifier and evaluate various ways to combine the two tasks, either as a single network with shared layers, or by feeding personality labels into the deception classifier. We show that including personality recognition improves the performance of deception detection on the Columbia X-Cultural Deception (CXD) corpus by more than 6% relative, achieving new state-of-the-art results on classification of phrase-like units in this corpus.

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

1. Introduction

Deception detection is a task of extreme interest and importance to numerous public and private sectors. It is also an extremely difficult task which even humans cannot perform with reliable accuracy. Automatic detection of deception from voice and language, which is noninvasive and relatively easy to collect and analyze, has been the object of much interest in the natural language processing (NLP) and speech community, and has yielded promising results.

Individual speaker differences such as personality play an important role in deception detection, adding considerably to its difficulty [1]. Enos et al. [2] also found that judges with different personalities perform differently when they detect deceit. We therefore hypothesize that personality scores may provide useful information to a deception classifier, helping to account for interpersonal differences in verbal and deceptive behavior.

In this study, we fuse a personality classification task with a deception classifier and evaluate various ways to combine the two tasks, either as a single network with shared layers, or by feeding personality labels into the deception classifier. We show that including personality recognition improves the performance of deception detection on the CXD corpus by more than 6% relative, achieving new state-of-the-art results on classification of phrase-like units in this corpus and demonstrating the capacity for personality information and multi-task learning to boost deception detection.

The remainder of this paper is structured as follows. In Section 2, we review previous work. A description of the data can be found in Section 3. In Section 4, we describe the feature sets and model architectures. Section 5 presents the experimental setup and results from the various models. Finally, we conclude and discuss future research directions in Section 6.

2. Related Work

Many researchers have explored the task of detecting deception from speech and language. Numerous lexical and acoustic-prosodic cues have been evaluated. Early work by Ekman et al. [3] and Streeter et al. [4] found pitch increases in deceptive speech. Linguistic Inquiry and Word Count (LIWC) categories were found to be useful in deception detection studies across five corpora, where subjects lied or told the truth about their opinions on controversial topics [5]. Other studies similarly report that deceptive statements can be distinguished from truthful statements using language-based cues, both verbal and nonverbal [6].

For speech-based deception detection, acoustic-prosodic features are used very often to identify the differences between deceptive and truthful speech, because pitch, energy, speaking rate and other stylistic factors may vary when speakers deceive. Hirschberg et al. [7] automatically extracted acoustic-prosodic and lexical features from Columbia-SRI-Colorado (CSC) corpus, the first cleanly recorded large-scale corpus of deceptive and non-deceptive speech. They achieved about 70% accuracy, and found that subject-dependent features were especially useful in capturing individual differences in deceptive behavior.

Interpersonal differences have been cited as a major obstacle to accurate deception detection [1]. Personality in particular has been associated with differences in deception *detection* behavior [2] However, there has been little research on using personality information to improve deception detection. Levitan et al. [8] found that including gender, native language, and personality scores along with acoustic-prosodic features improved classification accuracy on the Columbia Cross-Cultural Deception (CXD) Corpus [9], supporting the notion that deceptive behavior varies across different groups of people, and that including information about interpersonal differences can improve the performance of a deception classifier.

3. Data

The collection and design of Columbia X-Cultural Deception (CXD) Corpus analyzed here is described in more detail by [9, 10]. It contains within-subject deceptive and non-deceptive English speech 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. The data was collected using a fake resume paradigm, where pairs of speakers took turns interviewing their partner and being interviewed from a set of 24 biographical questions. Subjects were instructed to lie in their answers to a predetermined subset of the questions. Subjects were provided with financial incentive to lie effectively and judge deceptive statements correctly.

In addition to the high-level truth labels provided by the framework of the task, granular truth labels were reported by the participants as they spoke. While answering the biographical questions, each interviewee pressed the T or F key on a keyboard, labeling each utterance spoken as true or false. For example, in the middle of a deceptive statement about where they were born, a subject could include the truthful statement that their birthday was on a certain date. Similarly, truthful or further deceptive information could be included in the subject's responses to the interviewer's follow-up questions.

The keystrokes indicating granular truth labels were applied to speech segments according to the following alignment rule: [11]: a T or F label was assigned to a speech segment if a consistent label was retrieved (the interviewee pressed the corresponding key on the keyboard) between the start and end time of that segment, and eliminated otherwise.

Before the recorded interviews, subjects filled out the NEO-FFI (Five Factor) personality inventory [12], yielding scores for Openness to Experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). A median split of the Big Five scores is done to divide each of the big five personality groups into two classes, high and low.

Transcripts for the recordings were obtained using Amazon Mechanical Turk[1] (AMT), and the transcripts were force-aligned using Kaldi [13]. The speech was then automatically segmented into inter-pausal units (IPUs) using Praat, and transcripts were manually corrected. The subject key-presses were aligned with the speech as well.

In this study, the speech was segmented in two different ways: turn and IPU. An IPU is defined as speech from a single speaker separated by at least 50 ms silence, and a turn is defined as speech from a single speaker separated by at least 500 ms silence. Segments were eliminated if their duration is less than 0.05 seconds, resulting in average durations of 1.31 and 4.24 seconds for IPUs and turns, respectively. Finally, there are 79,632 and 30,368 IPU and turn level segments respectively, totaling 110,000 instances. Including instances from both levels of segmentation significantly increased the training size.

Table 1: *Segmentation Summary*

Duration	Avg (s)	Min (s)	Max (s)
IPU	1.31	0.05	21.76
Turn	4.24	0.06	115.41
Total	2.12	0.05	115.41
# Words	Avg (w)	Min (w)	Max (w)
IPU	4	1	47
Turn	11	1	387
Total	6	1	387

4. Methodology

4.1. Features

For our experiments, we use the feature sets described in [14, 15]: acoustic-prosodic low-level descriptor features (**LLD**); word category features from **LIWC** (Linguistic Inquiry and Word Count) [16]; and word scores for pleasantness, activation and imagery from the Dictionary of Affect in Language (**DAL**) [17]. We also use the Gensim library [18] to extract two sets of word embedding features (**WE**) using Google's pre-trained skip-gram vectors [19] and Stanford's pre-trained GloVe vectors [20].

In order to calculate the vector representation of a turn, we extract a 300-dimensional word vector for each word in the 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 signals, in addition to the higher-level psycholinguistic information represented by the LIWC and DAL features.

4.2. Deception models

Following Mendels et al. [11], we train three different models to predict deception: (1) a multilayer perceptron (MLP) trained using LIWC, DAL, LLD, and pretrained word embeddings, (2) a Long Short-Term Memory (LSTM) classifier using Stanford's pretrained GloVe vectors, and (3) a hybrid of the first two models. These models provide a baseline deception detection accuracy which we attempt to improve upon using personality prediction in Sections 4.3.2 and 4.3.1.

4.2.1. Multilayer perceptron

The multilayer perceptron (MLP) [21] is a simple feed-forward network using the sigmoid activation function. Our model has five fully-connected layers in a bottleneck configuration: (2048, 1024, 512, 1024, 2048) neurons per layer. The output layer consists of a single output neuron that uses the sigmoid function to calculate the probability of deception. During training, we add batch normalization [22] and a dropout layer [23] with 0.5 probability to each hidden layer.

4.2.2. Word Embedding and LSTM

We additionally experiment with feeding an instance's word embeddings into an LSTM (Long Short Term Memory) layer, well known for capturing sequential information [24, 25], to learn an instance-level representation.

For this model, which uses only the word embedding features as input, we also update the off-the-shelf word embeddings used in the MLP to better represent our data. We initialized a 300-dimensional word embedding layer with the Stanford off-the-shelf GloVe embeddings. We then trained the new model on our data, updating those initial weights. Since our corpus is relatively small, this takes advantage of the enormous corpora that were used to train the off-the-shelf embeddings, and adapts 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 instance 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 sigmoid function is then applied to the instance representation, outputting a probability estimation of the instance's deceptive status.

[1] https://www.mturk.com

4.2.3. *Hybrid Model*

A third model combines the previous two models by taking the output of the last hidden layer in the MLP model and concatenating it with the 256-dimensional output of the LSTM. The output of the concatenated layer is fed forward to a single output node that uses the sigmoid activation to predict the probability of deception.

4.3. Personality Recognition for Deception Detection

Motivated by [2, 26], we hypothesize that personality can improve deception detection. Therefore, we incorporate personality information into the deception detection models described in Section 4.2.

We evaluate two different ways to incorporate personality information into the deception classifier: (1) using multi-task learning to jointly predict both speaker personality and instance deception, and (2) feeding personality labels into the deception classifier's output layer.

4.3.1. *Multi-task learning*

The motivation behind multi-task learning — using a single classifier to predict two or more labels — is that robustness and accuracy may be improved by giving the classifier more than one task, since the tasks can influence each other through a shared representation.

For the multi-task MLP model, we add an output node for each personality trait to the last fully connected output layer with the sigmoid activation function (Figure 1). The output layer, which previously had a single node for predicting deception, now has five additional nodes, each of which predicts the probability that the instance speaker scored "high" on the corresponding Big Five personality trait. The output layers are similarly augmented for the LSTM and hybrid models.

In another variant of a multi-task learning model, the last hidden layer of the classifier feeds forward into five output sigmoid nodes that predict the Big Five personality traits. The output of the five personality classification nodes – five floating point numbers representing the probability of the speaker scoring "high" on each of the Big Five personality traits – are then concatenated back with the output hidden layer that preceded them, and the concatenation is fed forward to an output node for deception. Figure 2 shows how this looks in the MLP.

4.3.2. *Personality as a late feature*

For our second approach, instead of training a single classifier to predict both deception and personality, we feed personality labels into the deception classifier. The motivation is that the personality labels can act as features to inform the deception prediction. To reduce the chance that the impact of the personality features will be swallowed by the numerous other features, we introduce them to the classifier at a late stage, after the five hidden layers. This approach is the equivalent of Figure 2, without the links between the fully-connected layer and the personality nodes. That is, instead of the personality labels being predicted by the preceding layers, and influencing the weights of those layers through cross-entropy minimization, they are provided by an oracle: the gold standard labels self-reported by the speakers.

In a real-world system, these labels would be output by a separate or integrated personality classifier operating over the speech input. Since the model using the gold-standard labels gives an upper bound on how well such a model could perform, and its performance is exceeded by the multi-task models (Section 5), we do not further explore the potential of a model using personality labels predicted with various levels of accuracy.

Figure 1: *Diagram of multi-task learning MLP model (variant i).*

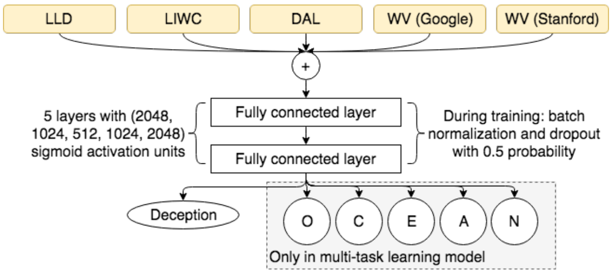
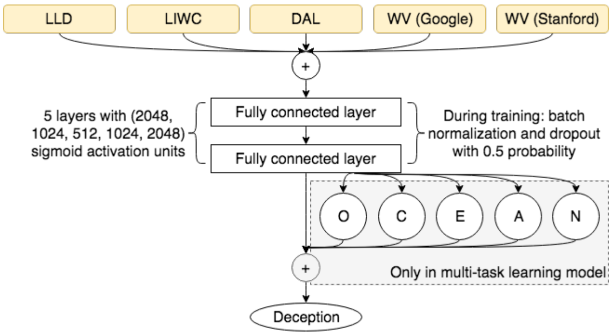

LLD
LIWC
DAL
WV (Google)
WV (Stanford)
+
5 layers with (2048, 1024, 512, 1024, 2048) sigmoid activation units
Fully connected layer
Fully connected layer
During training: batch normalization and dropout with 0.5 probability
Deception
O
C
E
A
N
Only in multi-task learning model

Figure 2: *Diagram of multi-task learning MLP model (variant ii).*


LLD
LIWC
DAL
WV (Google)
WV (Stanford)
+
5 layers with (2048, 1024, 512, 1024, 2048) sigmoid activation units
Fully connected layer
Fully connected layer
During training: batch normalization and dropout with 0.5 probability
O
C
E
A
N
+
Only in multi-task learning model
Deception

In both figures, the nodes labeled O, C, E, A, and N refer to the Big Five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

5. Results

In this section, we present various experimental results for deception detection with and without personality information. Table 2 shows the result of the deception-only models, Tables 3 and 4 show the results of the two multi-task models, and Table 5 shows the results of the model that includes personality labels as late features.

All models were trained using the Adam optimizer [27] with learning rate 0.001, decreasing at a rate of 50% for 100 epochs. Our data was split into train, validation and test sets of 83,600, 4,400, and 22,000 samples respectively. Since the classes (deceptive/nondeceptive) are unbalanced, we weighted each class based on the training set. All models were implemented using Keras [28].

As shown in Table 2, the best performance for deception detection without personality information is an F1 of 0.69, achieved by the MLP-LSTM hybrid model. This result can not be directly compared to the highest performance reported on this corpus so far, 0.64 [11], since that work predicted deception only at the level of IPUs, which are shorter and contain less information, and did not augment the training data with instances at the turn level. However, we can assume that our baseline

Table 2: *Deception detection without personality*

Model	Precision	Recall	F1
MLP	68.08	67.95	68.01
LSTM	65.64	66.08	65.78
Hybrid	69.43	69.46	69.45

Table 3: *Multi-task learning: variant (i)*

Model	Precision	Recall	F1
MLP	74.33	74.51	74.39
LSTM	64.61	65.56	64.40
Hybrid	69.42	69.67	69.51

Table 4: *Multi-task learning: variant (ii)*

Model	Precision	Recall	F1
MLP	74.37	74.67	74.38
LSTM	66.13	67.03	65.89
Hybrid	72.58	72.98	72.70

Table 5: *Deception detection using gold-standard personality as a late feature*

Model	Precision	Recall	F1
MLP	70.08	70.00	70.04
LSTM	65.09	65.74	65.22

deception-only model, which is based on the models presented in [11], achieves state-of-the-art level performance.

As shown in Tables 3 and 4, this performance can be improved by over 4% absolute by incorporating personality information into the model. Both approaches explored here, multi-task learning and adding personality as a feature, improved the F1 of the MLP from 0.68 to 0.744. While the first variant of multi-task learning did not significantly improve the hybrid model, the second variant increased its F1 to 0.727. Neither approach improved the LSTM model, with the first variant slightly reducing its F1.

We hypothesize that the difference between the MLP and LSTM models can be explained by the fact that the MLP has input from multiple feature sets. During training, the personality information – whether included as input or as output – can be used to adjust the weight matrix in the hidden layers to assign more weight to the features that are meaningful with respect to personality, performing a psychologically-informed form of feature selection that improves the deception detection performance. The LSTM model, on the other hand, uses sequential information from the instances' lexical content, and the concept of feature selection is less well-defined. Another possible explanation is that the personality information can interact meaningfully with the features from the acoustic and/or psycholinguistic signal, but are less informative with respect to the features from the instances' lexical content which are the only input to the LSTM.

Table 5 shows that including gold-standard personality labels improves the performance of the deception classification in the MLP, from 0.68 to 0.70. However, this model is outperformed by both multi-task learning models. This result is surprising, since the personality labels used in this model are the true ones reported by the instance speakers, while the personality information in the other models is predicted – perhaps inaccurately – from the instance features (a similar model predicting personality in this corpus reported an average of 60% accuracy [29]). This suggests that personality classification can assist the task of deception detection not only through the additional information of the speaker personality traits – captured by the 2% absolute improvement of this model – but also through the multi-task learning approach of influencing the shared layers towards a more useful and robust representation. An intriguing question for future work is whether this contribution is unique to personality classification, or whether a similar or added gain can be achieved by including additional classification tasks.

6. Conclusion and Future Work

In this paper, we present several approaches to combining the tasks of personality classification and deception detection. We compared the performance of MLP, LSTM, and hybrid models with multi-task learning and personality features. We found that both approaches to incorporating personality information into a deep deception classification model improved deception detection, adding to previous research indicating that deception detection can be improved by mitigating interpersonal differences. Multi-task learning performed better for deception detection than including personality features, suggesting a promising direction for improving deception detection.

Regarding individual model performance, we found that the MLP structure performed best in combination with multi-task learning, achieving the highest overall performance. Within-model performance improves by as much as 6% absolute when personality is added as a task, and the best model with personality (the multi-task MLP) performs 4% better than the best model without personality (the hybrid MLP-LSTM model).

In future work, we will explore the extension of these findings to other dataset and other related classification problems, such as adding gender and language classifiers to a deception classifier. We also see the potential of extending our framework to various analysis problems by embedding more paralinguistic and affective classifiers.

7. Acknowledgements

We thank Andrew Rosenberg and Michelle Levine. This work was partially funded by AFOSR FA9550-11-1-0120.

8. References

[1] A. Vrij, P. A. Granhag, and S. Porter, "Pitfalls and opportunities in nonverbal and verbal lie detection," *Psychological Science in the Public Interest*, vol. 11, no. 3, pp. 89–121, 2010.

[2] F. Enos, S. Benus, R. L. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg, "Personality factors in human deception detection: comparing human to machine performance." in *INTERSPEECH*, 2006.

[3] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Invited article: Face, voice, and body in detecting deceit," *Journal of nonverbal behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[4] L. A. Streeter, R. M. Krauss, V. Geller, C. Olson, and W. Apple, "Pitch changes during attempted deception." *Journal of personality and social psychology*, vol. 35, no. 5, p. 345, 1977.

[5] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[6] S. I. Levitan, G. An, M. Ma, R. Levitan, A. Rosenberg, and J. Hirschberg, "Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection." 2016.

[7] J. B. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis *et al.*, "Distinguishing deceptive from non-deceptive speech," 2005.

[8] 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 the Second Workshop on Computational Approaches to Deception Detection*, 2016, pp. 40–44.

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

[10] 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.

[11] G. Mendels, S. I. Levitan, K.-Z. Lee, and J. Hirschberg, "Hybrid acoustic-lexical deep learning approach for deception detection," *Proc. Interspeech 2017*, pp. 1472–1476, 2017.

[12] 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.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448–456.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

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

[25] 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.

[26] G. An, "Literature review for deception detection," 2015.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[29] G. An and R. Levitan, "Lexical and acoustic deep learning model for personality recognition," *Under submission*, 2018.