Supplemental Materials: Automatic Text-based Personality Recognition on Multiparty Dialogues Using Attentive Networks and Contextual Embeddings

Hang Jiang, [1] **Xianzhe Zhang,**[2] **Jinho D. Choi**[3]
[1]Symbolic Systems Program, Stanford University, CA 94305
[2]Department of Electrical Engineering, Stanford University, CA 94305
[3]Department of Computer Science, Emory University, Atlanta, GA 30322
hjian42@stanford.edu, xianzhez@stanford.edu, jinho.choi@emory.edu

Related Works

Big Five Theory

One of the most popular theories in personality variation is the Big Five (Norman 1963; Goldberg 1981). We adopt the following definitions of the five personality traits (John, Donahue, and Kentle 1991). Those definitions are below:

- **Agreeableness (AGR)**: forgiving, not demanding, warm, not stubborn, not show-off, sympathetic
- **Conscientiousness (CON)**: efficient, organized, not careless, thorough, not lazy, not impulsive
- **Extraversion (EXT)**: sociable, forceful, energetic, adventurous, enthusiastic, outgoing
- **Openness (OPN)**: curious, imaginative, artistic, wide interest, excitable, unconventional
- **Neuroticism (NEU)**: tense, irritable, not contented, shy, moody, not self-confident

Automatic Personality Recognition

Automatic text-based personality recognition, as an important topic in computational psycho-linguistics, focuses on determining oneś personality traits from text input. The Big Five Hypothesis is usually used for the task. Researchers have designed many linguistic markers, including lexical categories (Pennebaker and King 1999; Pennebaker, Mehl, and Niederhoffer 2003; Mehl, Gosling, and Pennebaker 2006; Fast and Funder 2008), n-grams (Oberlander and Gill 2006), and speech-act categories (Vogel and Vogel 1986), to analyze ones utterances and predict the personalities. Those features make personality recognition a feasible classification problem. Pennebaker et al. (Pennebaker, Francis, and Booth 2001) is one of the first researchers to extract Linguistic Inquiry and Word Count (LIWC) features (Mairesse and Walker 2006) from written essays to predict the five traits. Later studies focus on creating social media data (Golbeck et al. 2011; Schwartz et al. 2013; Park et al. 2015; Peng et al. 2015), designing linguistic features (Mairesse et al. 2007; Mohammad and Kiritchenko 2013; Celli et al. 2013), and conducting feature reduction (Tighe et al. 2016) to improve the field. Recently, Majumder et al. (2017) introduces Convolutional Neural Networks (CNN) with static word embeddings for the task on Essays dataset. This work outperforms the previous best model (Tighe et al. 2016) in CON, EXT, OPN, and NEU traits. However, no one has applied more robust attentive networks for classifying personality traits.

Attention-based Networks for Text Classification

Attention mechenism (Bahdanau, Cho, and Bengio 2014) was initially for neural machine translation. Recently, many NLP works have combined attention mechanism with Long Short-Term Memory (LSTM) and CNN networks for text classification (Zhou, Wan, and Xiao 2016; Chen et al. 2016; Shin et al. 2017). More complex architectures such as Hierachical attention networks (Yang et al. 2016) are proposed to both word and sentence level, allowing it to attend to more and less important information when constructing the document representation. These models achieve significant improvements over the baseline CNN and LSTM models. Therefore, we will experiment with a few attentive networks to improve the field of automatic personality recognition.

Contextual Word Embeddings

Word embedding is a popular feature learning technique in natural language processing (NLP). There are many training methods implemented by softwares such as Word2vec (Goldberg and Levy 2014), GloVe (Pennington, Socher, and Manning 2014), Gensim (Khosrovian, Pfahl, and Garousi 2008), and FastText (Joulin et al. 2016) to learn word representations. Recently, many efforts (ELMo, BERT, RoBERTa, XLNet) have been put into contextual embeddings. Contextual embeddings are assigned to each word in a sentence by language model instead of using a fixed word embeddings as we introduced before (Peters et al. 2018; Devlin et al. 2018; Liu et al. 2019; Yang et al. 2019). RoBERTa is the state-of-art text classifier for the General Language Understanding Evaluation (GLUE) benchmark. In our work, we decide to experiment with the base BERT model and its enhanced version, RoBERTa, for personality recognition. We also want to experiment with ELMo and XLNet in the future.

Dataset

FriendsPersona Dataset

Our new FriendsPersona dataset is developed upon the work of Chen and Choi, who built the conversational dataset from TV show (Chen and Choi 2016). Transcripts of the 10-season Friends TV show are formulated into clean JSON format. Each season has multiple episodes, and each episode has multiple scenes. Each scene is divided into utterances, and each utterance belongs to one speaker in the scene. One utterance has at least one sentence spoken by the speaker. This dataset has been widely used for NLP tasks such as emotion detection and character identification (Zahiri and Choi 2018; Choi and Chen 2018).

Example for MSF Algorithm To elaborate the *MainSpeakerFinder* algorithm, we make the following graph to illustrate how it works.

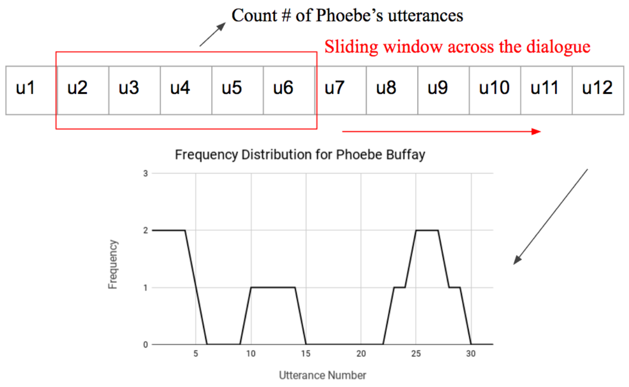
Count # of Phoebe's utterances
Sliding window across the dialogue
u1
u2
u3
u4
u5
u6
u7
u8
u9
u10
u11
u12
Frequency Distribution for Phoebe Buffay
Frequency
3
2
1
0
5
10
15
20
25
30
Utterance Number

Figure 1: An overview of the extraction algorithm. **U** indicates utterances.

For this algorithm (Fig 1), we first find a set of speakers in each scene. For each speaker, we use a sliding window to construct the frequency distribution graph in the scene. We pick the peaks in each frequency distribution graph if the peak frequency is bigger than or equal to a threshold like 2. Each peak is then used to identify the index range of consecutive sentences in which the speaker dominates temporarily. Each set of consecutive sentences extracted from the scene is called a sub-scene. In Fig 1, the two index ranges identified should be 0 to 6 and 22 to 30. In those two short conversations, Phoebe Buffay is a main character. We optimized the algorithm to get the maximum number of reliable conversations by setting the minimum utterance number of a conversation to be 5.

Annotation Workflow To annotate each sub-scene, an annotator needs to read the sub-scene first. At the end of the page, an annotator is asked to evaluate each of five personality traits on a -1 to 1 scale for one main speaker. 1 means the annotator agrees on that personality trait and -1 on the opposite of that personality trait. 0 means unclear or unknown.

Our annotation scheme is designed to be easy-to-understand and focuses on one speaker at a time. We expect an annotator to read the whole sub-scene to pick up the context and pay attention to the utterances from the target speaker in order to make valid judgments. We found out that our design indeed facilitates the process of the annotation.

Trait	Fleiss's Kappa	Average Pair-wise Kappa
AGR	23.50%	53.87%
CON	18.90%	54.34%
EXT	20.90%	57.81%
OPN	21.60%	56.12%
NEU	17.80%	52.46%
Average	20.54%	54.92%

Table 1: Inter-annotator agreement for each personality trait.

Inter-Annotator Agreement In the abstract, we report that average pair-wise kappa is 54.92% and Fleiss's kappa is 20.54% across five personality traits. We also include the individual agreement score for each trait (Table 1).

Experiments

Data Preprocessing

In Essays dataset, the average words per essay is long, 651 per essay. Therefore, we preprocess the data by filtering out all the punctuation marks and stop words. We also transform some expressions such as *I'll* and *I've* to *I will*, *I have*. After the text normalization step, we are able to shrink the size of the essays while preserving the most important information.

In `FriendsPersona` dataset, the average length per conversation is 160 words. We also apply the similar text normalization step as above. Besides, we replace speaker names at the beginning of each sentence with marks like 'speaker0' and 'speaker1'. Both datasets have binary class labels for each personality trait.

Analysis on Essays Dataset

We omit some experiments in the main abstract because of paper length limit, and we want to include these experiment results in the supplemental materials. In Table 2, there are a few experiments we conduct with LIWC linguistic features. We try Sequential Minimal Optimization (SMO), Logistics Regression (LR), and Multilayer Perceptron (MLP) with LIWC, and achieve comparable results as the state-of-art LIWC-based work (Tighe et al. 2016). Besides, we also include the results of Bidirectional LSTM (BLSTM) and TextCNN (Kim 2014) in the table. Their results are lower than those of ABLSTM (Attention-based BLSTM) and ABCNN (Attention-based CNN) respectively. This shows that attention mechanism is effective to boosting the performance of classification models. Since these models do not achieve results as good as some of other models, we choose to remove these models in the abstract and only include the most representative models (ABCNN, ABLSTM, HAN, BERT, RoBERTa).

Unlike the previous work (Majumder et al. 2017), We use FastText embeddings for baseline attentive models instead of GloVe embeddings in the experiment because FastText is more superior so that it composes word embeddings for out-of-vocabulary (OOV) words with char n-grams instead of using an unknown token. Our FastText embeddings are trained on a corpus combining New York Times, Wikipedia dump, the Amazon Book Reviews, and the *Friends* transcripts.

Our code is published on Github[1]. The full list of models we cite and experiment are included below:

- **LIWC** represents the best results achieved by classic machine learning models trained on Linguistic Inquiry and Word Count (LIWC) features with feature reduction technique (Tighe et al. 2016). This model has the state-of-art result on AGR trait.
- **HCNN** represents Hierarchical Convolutional Neural Networks (Majumder et al. 2017). This model has the state-of-art results on CON, EXT, OPN, NEU traits.
- **LIWC + SMO** represents Sequential Minimal Optimization trained on LIWC linguistic features.
- **LIWC + LogisticsRegression** represents Logistics Regression trained on LIWC linguistic features.
- **LIWC + MLP** represents Multilayer Perceptron trained on LIWC linguistic features.
- **MLP** represents Multilayer Perceptron with FastText embeddings.
- **TextCNN** represents text Convolutional Neural Networks (CNN) (Kim 2014) with FastText embeddings.
- **ABCNN** represents TextCNN with global attention mechanism (Bahdanau, Cho, and Bengio 2014) with FastText embeddings.
- **BLSTM** represents Bidirectional Long Short-Term Memory Networks (BLSTM) (Hochreiter and Schmidhuber 1997; Schuster and Paliwal 1997) with FastText embeddings.
- **ABLSTM** represents Bidirectional LSTM with global attention mechanism (Bahdanau, Cho, and Bengio 2014) with FastText embeddings.
- **HAN** represents hierarchical attention networks (Yang et al. 2016) with FastText embeddings.
- **BERT** represents the classification model with Pre-training of deep bidirectional transformers (BERT) (Devlin et al. 2018). We use the pre-trained base BERT model[2] released by Google.
- **RoBERTa** represents the classification model with Robustly Optimized BERT model (Liu et al. 2019). We use the pre-trained base RoBERTa model[3] released by Facebook.

Analysis on FriendsPersona Dataset

On the FriendsPersona dataset, we only experiment with the most representative models (ABCNN, ABLSTM, HAN, BERT, RoBERTa) because the omitted models (LIWC-based models, MLP, CNN, BLSTM) in the previous experiments are not performing as well as the selected models. Our goal is to establish a baseline on our dataset with the best models for the future works.

Models	**AGR**	**CON**	**EXT**	**OPN**	**NEU**
Majority Baseline	53.08	50.81	51.74	51.54	50.04
LIWC (Tighe et al. 2016)	**57.50**	56.00	55.70	61.95	58.30
HCNN (Majumder et al. 2017)	56.71	**57.30**	**58.09**	**62.68**	**59.38**
LIWC + SMO*	56.60	54.98	54.34	61.18	57.33
LIWC + LogisticsRegression*	57.17	55.02	53.36	60.82	58.10
LIWC + MLP*	57.90	56.62	55.96	57.62	56.72
MLP*	55.51	58.59	56.69	59.44	56.93
TextCNN*	57.38	57.74	56.28	63.49	57.09
ABCNN	57.82	**60.13**	58.75	63.65	58.51
BLSTM*	56.64	57.83	59.17	63.02	57.69
ABLSTM	58.85	59.55	59.32	63.99	59.56
HAN	57.62	59.32	59.77	63.61	58.75
BERT	58.10	57.69	59.12	61.17	59.20
RoBERTa	**59.72**	58.55	**60.62**	**65.86**	**61.07**

Table 2: The performance of models in accuracy on Essays. * represents experiments missing in the abstract.

The main limitation of our models is that they do not make full use of the dialogue structure to leverage information between the target speaker and other speakers. That is why we experiment with 3 data formats to force the models to distinguish the target speaker's utterances and the context information. For ABCNN, ABLSTM, BERT, and RoBERTa models, we still treat the the whole dialogue text as one long sentence for classification. HAN is able to distinguish each utterance using its hierarchical structure. This explains why HAN does comparably better on the dialogue dataset than on the monologue dataset. However, it is still not designed to model the interactions between the target speaker and other speakers.

In the future, we want to create a novel attentive model that interprets the target speaker's utterance by querying the context information to predict the speaker's personality traits.

Acknowledgments

We want to acknowledge the support of EmoryNLP in funding our work. Besides, we want to thank James W. Pennebaker for sharing the Essays dataset for our research.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Celli, F.; Pianesi, F.; Stillwell, D.; and Kosinski, M. 2013. Workshop on computational personality recognition: Shared task. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Chen, Y.-H., and Choi, J. D. 2016. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In *SIGDIAL Conference*, 90–100.

Chen, H.; Sun, M.; Tu, C.; Lin, Y.; and Liu, Z. 2016. Neural sentiment classification with user and product attention. In

[1] https://github.com/emoryjianghang/automatic-personality-prediction

[2] https://github.com/google-research/bert

[3] https://github.com/pytorch/fairseq/tree/master/examples/roberta

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1650–1659.

Choi, J. D., and Chen, H. Y. 2018. Semeval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 57–64.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fast, L. A., and Funder, D. C. 2008. Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology* 94(2):334.

Golbeck, J.; Robles, C.; Edmondson, M.; and Turner, K. 2011. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, 149–156. IEEE.

Goldberg, Y., and Levy, O. 2014. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Goldberg, L. R. 1981. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology* 2(1):141–165.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

John, O. P.; Donahue, E. M.; and Kentle, R. L. 1991. The big five inventoryversions 4a and 54.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Khosrovian, K.; Pfahl, D.; and Garousi, V. 2008. Gensim 2.0: a customizable process simulation model for software process evaluation. In *International Conference on Software Process*, 294–306. Springer.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mairesse, F., and Walker, M. 2006. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 85–88. Association for Computational Linguistics.

Mairesse, F.; Walker, M. A.; Mehl, M. R.; and Moore, R. K. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30:457–500.

Majumder, N.; Poria, S.; Gelbukh, A.; and Cambria, E. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems* 32(2):74.

Mehl, M. R.; Gosling, S. D.; and Pennebaker, J. W. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology* 90(5):862.

Mohammad, S. M., and Kiritchenko, S. 2013. Using nuances of emotion to identify personality. *Proceedings of ICWSM*.

Norman, W. T. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology* 66(6):574.

Oberlander, J., and Gill, A. J. 2006. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes* 42(3):239–270.

Park, G.; Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Kosinski, M.; Stillwell, D. J.; Ungar, L. H.; and Seligman, M. E. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology* 108(6):934.

Peng, K.-H.; Liou, L.-H.; Chang, C.-S.; and Lee, D.-S. 2015. Predicting personality traits of chinese users based on facebook wall posts. In *Wireless and Optical Communication Conference (WOCC), 2015 24th*, 9–14. IEEE.

Pennebaker, J. W., and King, L. A. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77(6):1296.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54(1):547–577.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

Shin, B.; Chokshi, F. H.; Lee, T.; and Choi, J. D. 2017. Classification of radiology reports using neural attention models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 4363–4370. IEEE.

Tighe, E. P.; Ureta, J. C.; Pollo, B. A. L.; Cheng, C. K.; and de Dios Bulos, R. 2016. Personality trait classification of essays with the application of feature reduction. In *SAAIP@ IJCAI*, 22–28.

Vogel, K., and Vogel, S. 1986. L'interlangue et la personnalite de l'apprenant. *IRAL: International Review of Applied Linguistics in Language Teaching* 24(1):48.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zahiri, S. M., and Choi, J. D. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, X.; Wan, X.; and Xiao, J. 2016. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 247–256.