Prevalence of Personal Attributes Found in Twitter Posts

Take Yo* Kazutoshi Sasahara†

Abstract

This study aims to investigate the extent to which personal attributes can be predicted from social media posts. We focused on finding the baseline accuracy of using social media text data to predict personal attributes, and identifying the mechanism and crucial elements of the prediction process.The results exhibited approximately 70% accuracy when inferring most of the 12 personal attributes. The study also exhibited a strong influence of nouns compared to other part of speech (POS), a strong effect of shuffling the corpus.

1 Introduction

Computational social science (CSS) is not only a quantitative reinforcement of existing social sciences, but an unprecedented extension of the range of social sciences at a methodological level. CSS has become a key tool for understanding the complex relationship between human behaviors and social phenomena in the information age.

As an important field of CSS, the inference of personal attributes from social data is increasingly studied in both academia and industry. It can be applied to a wide range of areas, including basic research in social science and applications for information recommendation and social media marketing [13, 14].

Schwartz et al. analyzed words, phrases, and topics from Facebook posts. Combined with personality tests, they observed close relationships among language use and personality, gender, and age [17]. Kosinski et al. demonstrated that even simple algorithms could predict personal attributes on the bias of the patterns of Facebook's "likes," an indicator of people's preferences [11]. Liu and Zhu demonstrated that the Big Five factors in human personality (i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism) could be predicted from text posts on the Chinese microblogging platform Weibo [15]. IBM has also developed a service called Personality Insights that predicts personality traits, such as the Big Five factors, needs, and values [3].

*Nagoya University, yotake1987@nagoya-u.jp

†Nagoya University, JST PRESTO, sasahara@nagoya-u.jp

Although a significant amount of studies have been done on this topic, little is known about how to computationally infer personal attributes from social data, and no versatile algorithm has yet been established. The aim of this study is to investigate the extent to which personal attributes can be predicted from social media posts.

2 Data and Method

2.1 Data Collection

We recruited 703 participants to answer a questionnaire. Each participant answered 12 questions about their personal attributes. Each participant also agreed to share their tweets for research purposes. For each Twitter account, we gathered all posts from its timeline through Twitter API. Then, we picked out all posts from 650 active Twitter accounts with more than 2000 posts, including tweets, retweets, and replies. As a result, we built a dataset containing 1,950,00 tweets, which we split into a training set of 1,625,000 tweets and a test set of 325,000 tweets.

2.2 Data Processing

Data processing was performed in the following steps. First, the Japanese tweets we collected were segmented into words using the Japanese morphological analysis tool Mecab [10] with the Japanese dictionary NEologd [5]. Segmented tweets shorter than four words in length were considered less informative and were therefore deleted. Then, word2vec [16] was used as a word embedding method to create a dictionary of word vectors from the segmented tweets for training. We used a Skip-gram model with a window size of 5 and 20 iteration times for word2vec implemented in the machine learning framework Chainer [6]. Although doc2vec [12] is often used for vectorization of sentences, it is unlikely to work for short sentences, such as tweets. Instead of doc2vec, we used the method of averaging word vectors to obtain tweet vectors.

Tweets vectors were constructed based on data from the training accounts by referencing the dictionary of word vectors created previously. The same procedure was applied to the data from the test accounts. Some words existed in the tweets from the test accounts but did not exist in the dictionary of word vectors. We therefore did not use these words for constructing tweet vectors.

Since a tweet consist of 140 characters or less, a single tweet may not convey enough personal information for analysis, but a collection of multiple tweets might be a more effective unit for inferring personal attributes. Thus, we used a group of tweets or “tweet blocks” as inputs for machine learning and tweet block vectors were constructed by averaging the tweet vectors.

2.3 Algorithms

Using machine learning algorithms, we trained and tested models based on single tweets ($L = 1$) or tweet blocks ($L > 1$) obtained from the above-mentioned processing method. We used scikit-learn [8] for Linear Support Vector Classification (Linear SVC) [9], K-Neighbors [4], AdaBoost [1], and Random Forest [7]. The best parameters for these algorithms were selected using 6-fold cross-validation. Furthermore, we used Chainer for deep learning, in which the number of middle layers is 2, the number of nodes for each layer is 50 and 25, all the types of activation functions are sigmoid and the optimizer is Adam with learning rate 0.001. These parameters were optimized through repeated trials.

2.4 Analysis of Prediction Accuracy

We focused on the influence of factors that contribute to the prediction performance. We analyzed each factors individually: parts of speech (POS) and the order of the sequence of tweets.

To analyze the influence of the POS of the corpus, we applied the analysis, which consisted of filtering out all words except for the POS from each tweet before the vectorization. We then compared the accuracy with the baseline accuracy.

To analyze the influence of the order of the sequence of tweets, we followed the sequence of upload timestamp not only in collecting user's timelines but also generating tweet blocks. To investigate the influence of the time sequence of tweets, we regenerated tweet blocks after shuffling all the tweets and then compared the accuracy of prediction of personal attributes to the original data.

3 Results

3.1 Prediction of 12 Personal Attributes

Fig 1 shows the accuracy of different learning algorithms for predicting 12 personal attributes in the combination of N=100, L=50. All the algorithms exhibited approximately 70% accuracy when inferring all personal attributes except "attitude toward alcohol," "style of pet-lover," and "preferred method of chat." Regarding "gender," "has children or not," "age group," and "has more than 150 (i.e. Dumber's number [2]) Facebook friends or not," accuracy reached over 75%. The results also showed that deep learning exhibited better performance and higher accuracy compared to the other three algorithms.

3.2 Effects of Tweet Block Size on Prediction Accuracy

Because the results described in Fig 1 showed better performance in deep learning than did the three other algorithms, we applied deep learning to further investigate the effects of tweet block size (L) on prediction accuracy. Fig 2 shows accuracy as a function of different L under the condition of the same N (100) and

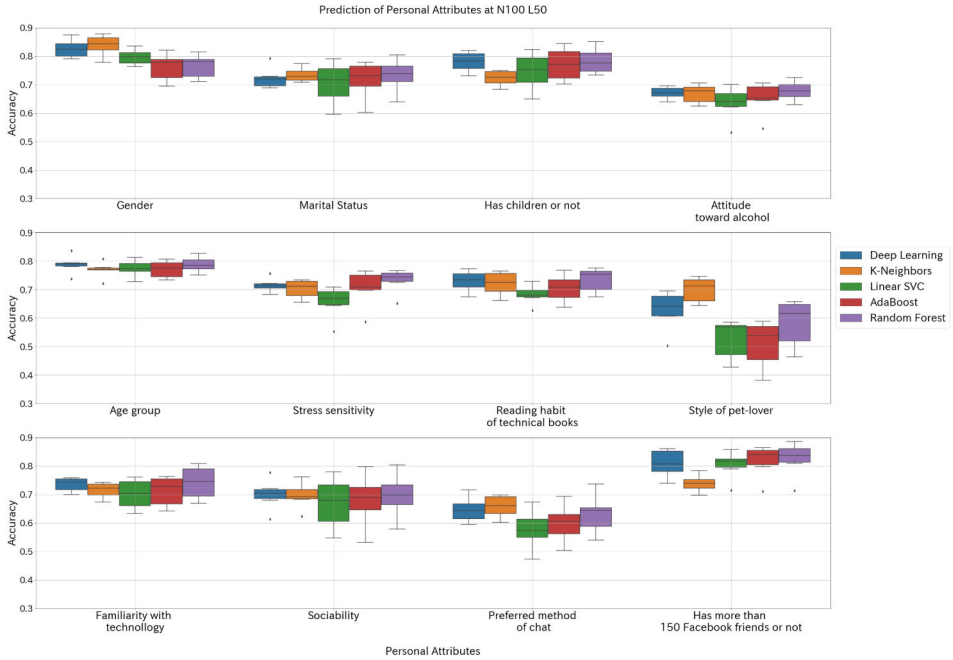

Figure 1: Comparison of different algorithms in prediction accuracy for 12 personal attributes. All results show the mean value and standard deviation of six fold cross validations.

algorithm. Prediction accuracy for tweet blocks (L>1) is higher than for single tweets (L=1) in all 12 personal attributes. However, in most cases, after reaching the value of L=50, prediction accuracy did not significantly increase.

3.3 Analysis of Related Factors in the Prediction Process

When analyzing the influence of the POS, we first individually extracted the nouns, adjectives and verbs from each tweet before vectorization, using the word-embedding model generated from the original corpus). Then we compared the accuracy with the baseline accuracy shown in Fig 3. The results indicate that in the word embedding of the original corpus, nouns (occupied 48.96% of the whole original bag of words) shows the same level of high prediction accuracy as the original, compared to adjectives and verbs.

Fig 4 shows the results of the comparison between the shuffled corpus and the original corpus. Compared to the original unshuffled corpus, which followed the sequence of the upload time stamp, the result of the shuffled version was much more unstable. However, this difference has not been proven to be statistically significant (t-test, $P > 0.05$).

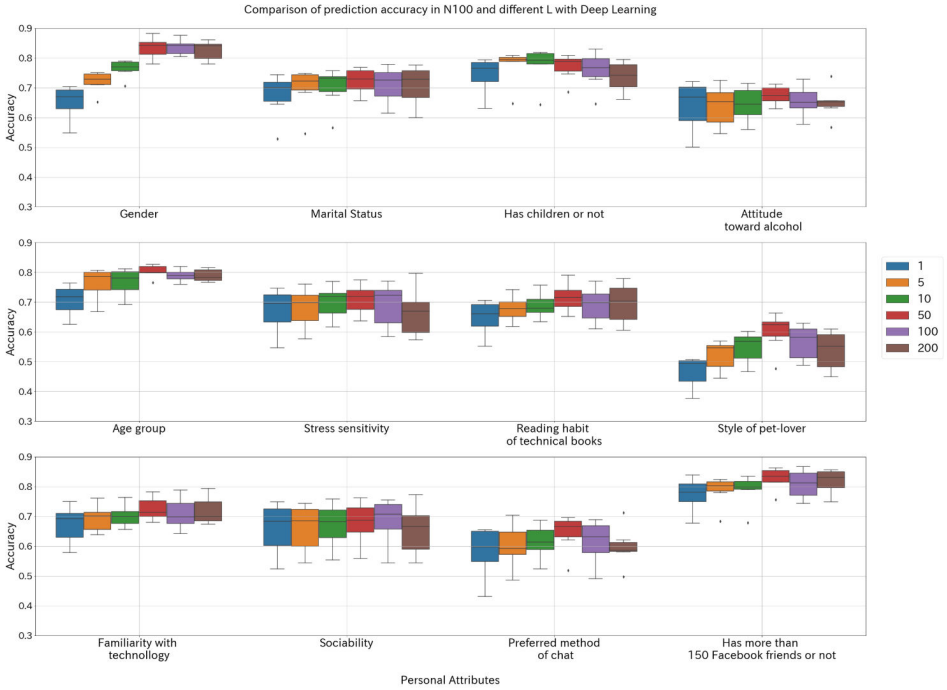

Figure 2: Comparison of different tweet block sizes (L) in prediction accuracy for 12 personal attributes with deep learning. All results show the mean value and standard deviation of six fold cross validations.

4 Discussion

Besides the basic personal attributes ("gender" and "age group"), "has children or not" and "has more than 150 Facebook friends or not" reached over 75% accuracy. The results of the Twitter data set can even predict some information from other SNS platforms, which implies the possibility of discovering personal attributes that exist in various SNS platforms.

Results showed that tweet blocks showed significantly higher accuracy for all prediction tasks compared to single tweets. Furthermore, we found that in some personal attributes (e.g., "stress sensitivity," "reading habit of technical books," "style of pet-lover,"), a larger value of L even decreased prediction accuracy. It also suggests that the optimal value of L may vary with the kind of personal attributes, just like with other parameters of the prediction.

In the analysis of related factors, the results from Fig 3 suggest that the high ratio of nouns in the corpus influences prediction accuracy, while the high ratio of embedding hyperlinks does not significantly affect prediction accuracy. Based on these results, we could noticeably reduce the size of the corpus (to 48.28% of the original) to obtain almost the same level of performance compared to the original

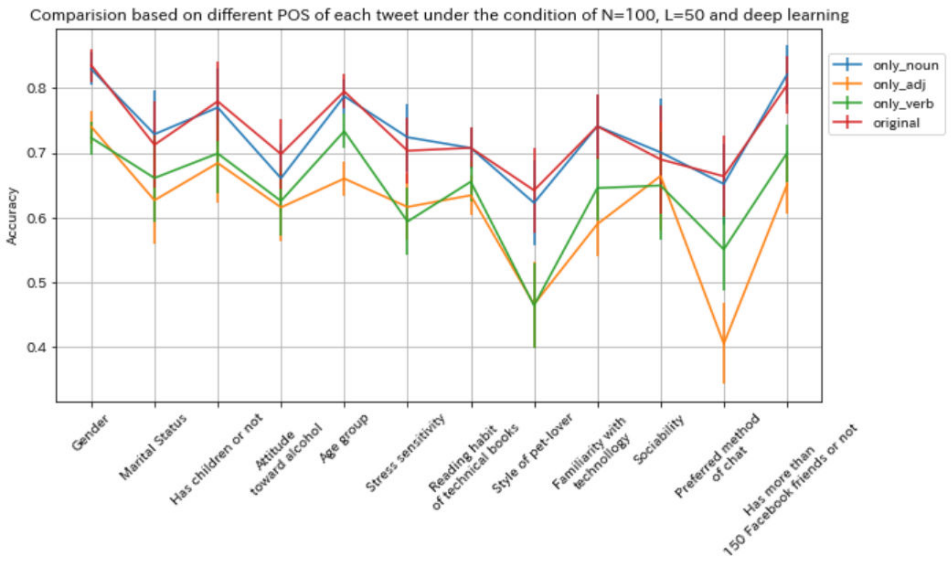

Figure 3: Comparison of prediction accuracy based on different POS of each tweet under the condition of N=100, L=50 and deep learning for 12 personal attributes. All results are shown by the mean value and standard deviation of six fold cross validations.

corpus.

Despite the near lack of difference between the shuffled corpus and the unshuffled corpus, the results shown in Fig 4 still suggest that there may be some unstable influence from the shuffled processing. Thus, the true effects need to be verified in the future.

Acknowledgement

This research was supported by JSPS/MEXT KAKENHI Grant Number JP17H06383 in #4903, JST PRESTO Grant Number JPMJPR16D6, and JST CREST Grant Number JPMJCR17A4.

References

[1] Adaboost. https://en.wikipedia.org/wiki/AdaBoost.

[2] Dunbar's number. https://en.wikipedia.org/wiki/Dunbar%27s_number.

[3] IBM Watson Personality Insights. https://www.ibm.com/watson/services/personality-insights/.

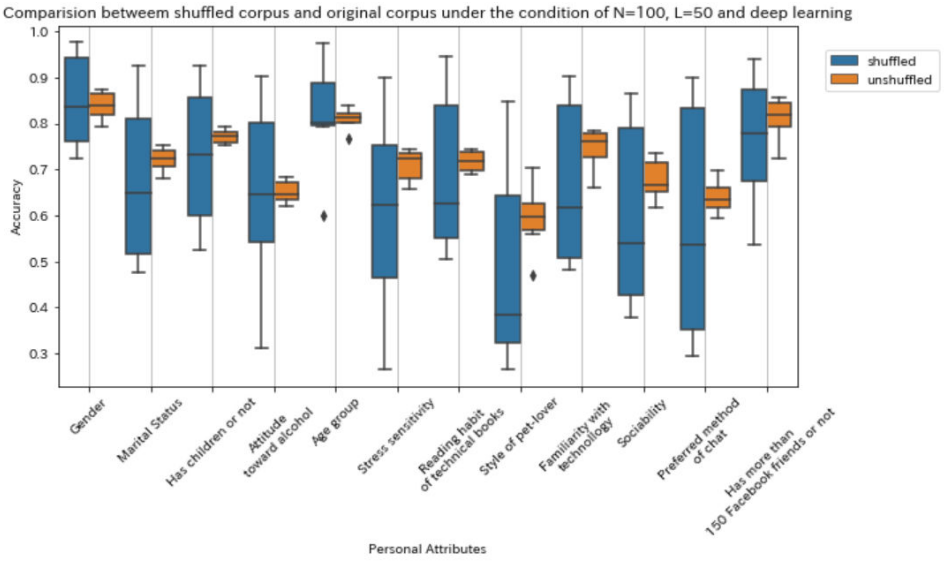

Figure 4: Comparison of prediction accuracy between the shuffled corpus and the original corpus under the condition of N=100, L=50 and deep learning for 12 personal attributes. All results are shown by the distributions of results from six fold cross validations.

[4] K-nearest neighbors. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

[5] mecab-ipadic-neologd: Neologism dictionary for mecab. https://github.com/neologd/mecab-ipadic-neologd.

[6] A powerful, flexible, and intuitive framework for neural networks. https://chainer.org.

[7] Random forest. https://en.wikipedia.org/wiki/Random_forest.

[8] scikit-learn: Machine learning in python. http://scikit-learn.org/stable/.

[9] Support vector machine. https://en.wikipedia.org/wiki/Support-vector_machine.

[10] Mecab: Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net, 2005.

[11] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[12] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

[13] Brent Smith Linden, Greg and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, (1):76–80, 2003.

[14] et al. Lipsman, Andrew. The power of "like": How brands reach (and influence) fans through social-media marketing. *Journal of Advertising research*, (52):40–52, 2012.

[15] Xiaoqian Liu and Tingshao Zhu. Deep learning for constructing microblog behavior representation to identify social media user's personality. *PeerJ Computer Science*, 2:e81, 2016.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.

[17] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791, 2013.