

Inference of Personal Attributes from Tweets Using Machine Learning

Take Yo*

* *Department of Complex Systems Science,
Graduate School of Informatics, Nagoya University,
Nagoya, Japan
Email: yotake1987@nagoya-u.jp*

Kazutoshi Sasahara*†

**Department of Complex Systems Science,
Graduate School of Informatics, Nagoya University,
Nagoya, Japan
†JST, PRESTO, Kawaguchi, Japan
Correspondence Email: sasahara@nagoya-u.jp*

Abstract—Using machine learning algorithms, including deep learning, we studied the prediction of personal attributes from the text of tweets, such as gender, occupation, and age groups. We applied word2vec to construct word vectors, which were then used to vectorize tweet blocks. The resulting tweet vectors were used as inputs for training models, and the prediction accuracy of those models was examined as a function of the dimension of the tweet vectors and the size of the tweet blocks. The results showed that the machine learning algorithms could predict the three personal attributes of interest with 60–70% accuracy.

Keywords—computational social science; deep learning; machine learning; personal attribute; social media.

I. INTRODUCTION

Social media was developed for people who desire diverse communication paradigms. Currently, social media plays the role of a hub for social information, a platform for exchanging opinions, and a place for researchers to observe digital traces of human behavior. In addition to the availability of big data and improvements in computing capability with GPGPU, machine learning techniques, including deep learning, are more practical for image, sound, and natural language processing. They are becoming important tools for promoting the social sciences and constructing social information infrastructures [1]–[3]. Based on this background, a new interdisciplinary science called computational social science has emerged and it has been actively investigated in recent years [4], [5].

According to the official announcement from Twitter, as of April 2016, there are approximately 310 million monthly active accounts worldwide, and the number of monthly views of tweets with embedded photos and videos has reached 1 billion [6]. Although the increase in the number of active accounts is slowing year after year for social media competitors (such as Facebook and Instagram), there are about 40 million monthly active Twitter accounts in Japan, as of September 2016. This indicates that Twitter is still a popular social media platform in Japan [7]. Unlike Facebook, which requires

permissions for social networking, one can follow both friends and others of interest without permissions. After that, friends can share any contents online. This follow mechanism enables “loose social relationships” that encourage diverse communications. Besides this, Twitter allows us to collect a large amount of social data via API [8]. Thus, Twitter has frequently been used as a data source by computational social science researchers (e.g., [9]–[12]).

The prediction of personal attributes based on social data has become a major research theme in recent years. Previous research has examined the relationships between personal attributes and behaviors. For example, how people talk and write are known to be associated with various personal attributes, such as educational background and growth environment [13], [14]. However, previous research has been limited by the availability of data. In the age of social media, people spontaneously post and share linguistic expressions online, and this information can be used to infer personal attributes [5]. Although a significant amount of research has 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 research is to investigate the extent to which personal attributes can be predicted only from texts in social media posts. This results can give us baseline data; we can further increase the prediction accuracy by adding other features in addition to texts. In this study, we used a large dataset from Twitter, transferred tweets to vectors using a word embedding method, and then predicted gender (male or female), occupation (10 different jobs), and age groups (whether he/she was born before 1980, indicating “digital native” or “digital immigrant”) based on those tweet vectors using five different machine learning algorithms. In this paper, we report two preliminary results.

II. RELATED RESEARCH

With the development of information technology, the inference of personal attributes based on social data has been actively studied for applicability.

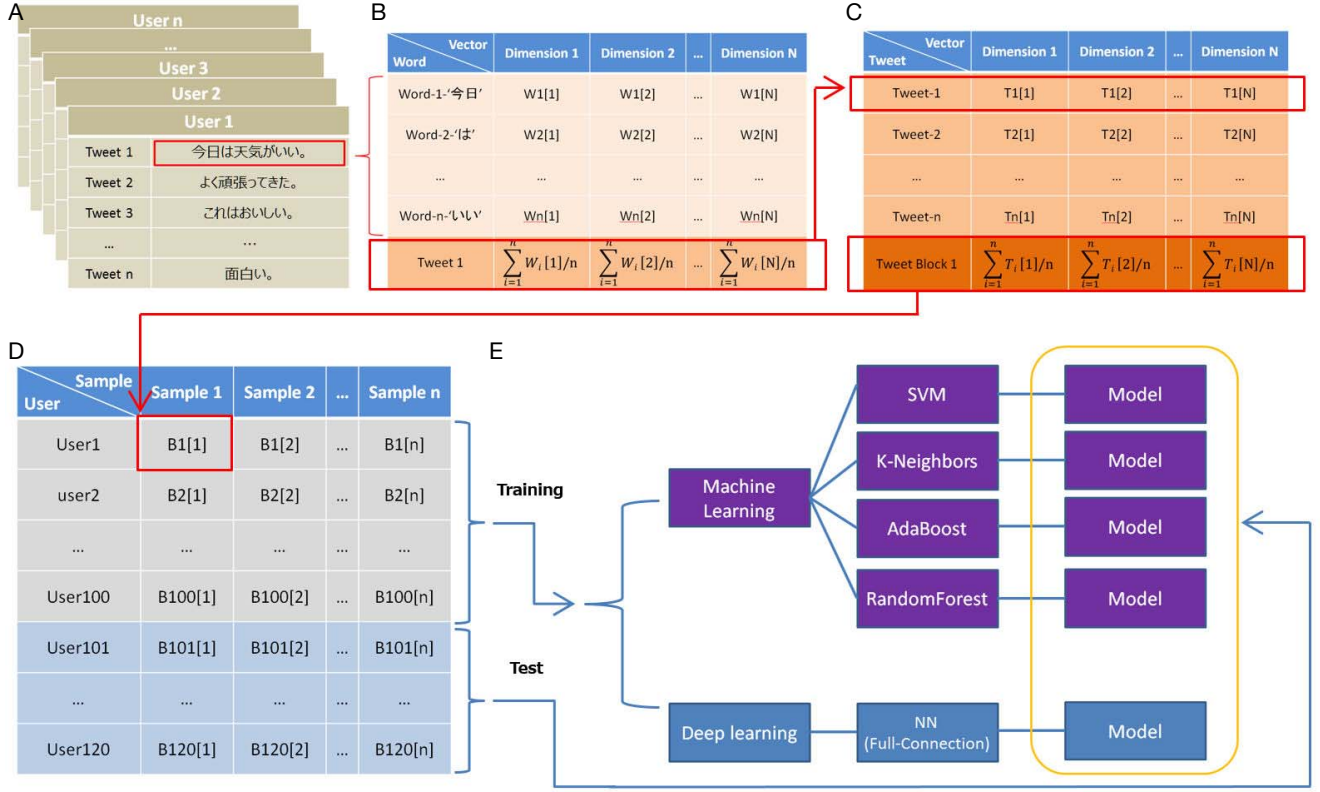


Figure 1. Schematic diagram of data processing and machine learning

Sloan et al. showed that demographic information (gender, language, location, age, occupation, and social class) could be accurately extracted from the profile descriptions of Twitter users using natural language processing (NLP) [15], [16]. 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 can predict personal attributes on the basis of the patterns of Facebook’s “likes,” an indicator of peoples’ preferences [18]. Wang et al. applied various deep learning algorithms to extract information from tweets, profile images and posted pictures and predicted their political intonation [19]. 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 [20]. IBM has also developed a service called Personality Insights which

predicts personality traits including the Big Five factors, needs, and values [21].

The inference of personal attributes from social data is increasingly studied in academia and industry because 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.

III. METHOD

A. Data Collection

To construct models that predict gender, occupation, and age groups, we collected tweets from Japanese accounts via. Twitter API [8]. We manually selected 120 active Twitter accounts whose posts numbered more than 3000 at the time of crawling and for which genders and occupations could be identified from the personal Twitter profiles or reliable information sources, such as Wikipedia. We then assigned age group labels to these accounts using reliable information sources as much as possible, because age information is less readily available

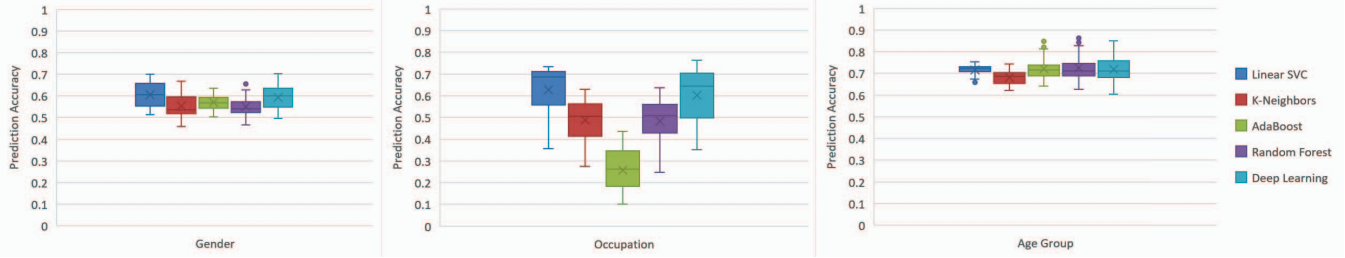


Figure 2. Comparison of different algorithms in prediction accuracy for three attributes

compared to gender and occupation. Among the 120 accounts, we used tweets from 100 accounts for training and tweets from the remaining 20 accounts for testing.

In the 120 accounts, the number of males and females was the same. The number of accounts belonging to each of 10 types of jobs (politician, entertainer, cartoonist, entrepreneur, scholar, journalist, writer, musician, athlete, IT engineer) was also the same. The age groups were divided into “digital natives” born after 1980, “digital immigrants” born before 1980, and “unknowns” who were not identified via internet search. In the training data from 100 accounts, there were 50 digital natives, 41 digital immigrants, and 9 unknowns. In the test data from 20 accounts, there were 11 digital natives, and 9 digital immigrants. Thus, the number of people in age group was approximately the same. Next, we crawled user timelines, which included retweets and replies from each account as much as possible (Fig. 1A). This resulted in 314,382 tweets from the training accounts and 64,027 tweets from the test accounts.

B. 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 [22] with the Japanese dictionary NEologd [23]. Segmented tweets shorter than four words in length were considered less informative and deleted. As a result, there remained 312,169 tweets for training and 63,454 tweets for testing. In this research, word2vec [24] was used as a word embedding method to create a dictionary of word vectors from the segmented tweets for training (30,5491 different words in 11,308,535 total words). 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 [25]. Although doc2vec [26] is often used for vectorization of sentences, it is unlikely to work for short sentences, such as tweets. Thus, instead of doc2vec, we used the method of averaging word vectors in order to obtain tweet vectors, as

shown below (Fig. 1B):

$$T = \sum_{i=1}^n \frac{w_i}{n},$$

where T is a tweet vector, W_i is the vector of i th word in a given tweet, and n is the number of words in the tweet.

Tweets vectors were constructed based on data from 100 training accounts by referencing the dictionary of word vectors created previously (Fig. 1D). The same procedure was applied to data from 20 test accounts. Some words exist in tweets from 20 test accounts but did not exist in the dictionary of word vectors. We did not use these words for constructing tweet vectors. The number of unused words was only 4% of the total words in the tweets from 20 test accounts.

Since a tweet consist of 140 characters or less, a single tweet may not convey enough personal information, 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 used (Fig. 1C and D).

C. Machine Learning Algorithms

Using machine learning algorithms, we trained and tested models based on single tweets ($L=1$) or either tweet blocks ($L > 1$) obtained from the above-mentioned processing method (Fig. 1E). We used scikit-learn [27] for Linear Support Vector Classification (Linear SVC), K-Neighbors, AdaBoost, and Random Forest. The best parameters for these algorithms were selected using 10-fold cross-validation. Furthermore, we used Chainer for deep learning, in which a full connection neural network was chosen and parameters such as the number of middle layers, the number of nodes for each layer, learning rate, and the types of activation functions were optimized through repeated trials.

IV. RESULTS

We examined the effects of the word embedding dimension (N) on word2vec, tweet block size (L), and

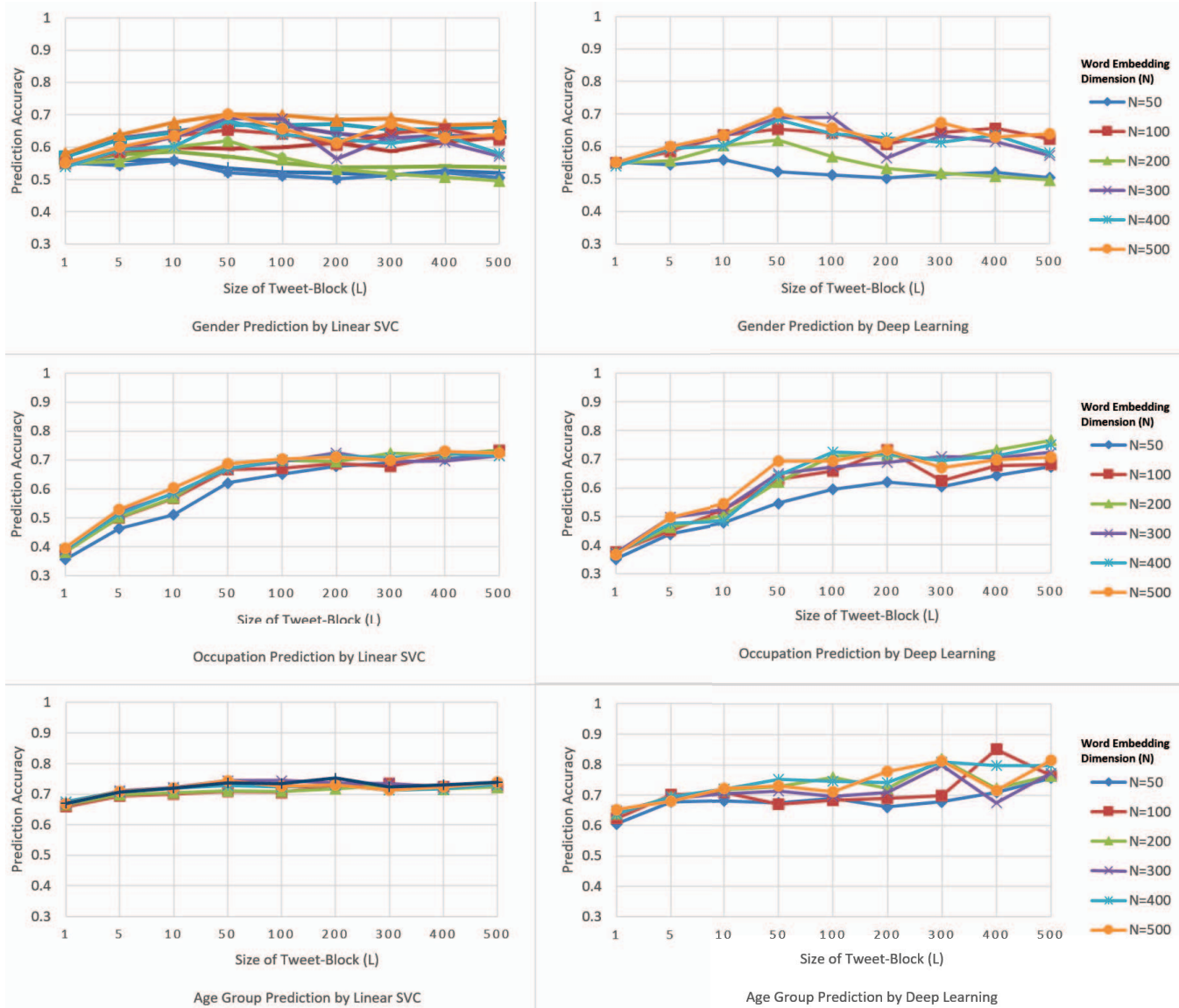


Figure 3. Comparison of prediction accuracy between linear SVC and deep learning

different algorithms in terms of the prediction accuracy of three personal attributes: gender, occupation, and age groups.

A. Effects of Word Embedding Dimension and Tweet Block Size on Prediction Accuracy

Figure 2 shows the accuracy of different learning algorithms for three kinds of prediction tasks in all combinations of N and L . All the algorithms exhibited approximately 70% accuracy with respect to inferring age groups for (digital natives and digital immigrants), indicating that age groups can be more easily predicted from social data than the other two attributes. The result also showed that Linear SVC and deep learning ex-

hibit stable performance and higher accuracy compared with the two other algorithms.

Figure 3 shows the accuracy distributions of Linear SVC and deep learning for each attributes for different values of N and L . For gender prediction, both Linear SVC and deep learning achieved approximately 70% accuracy at $N = 500$ and $L = 50$ (F-Score = 0.688 and AUC-Score = 0.701 in Linear SVC; F-Score = 0.702 and AUC-Score = 0.706 in deep learning; male (positive) / female (negative)). For occupation prediction, Linear SVC achieved approximately 70% accuracy at $N = 500$ and $L = 100$, and deep learning achieved a comparable level of accuracy at $N = 200$ and $L = 100$. For age

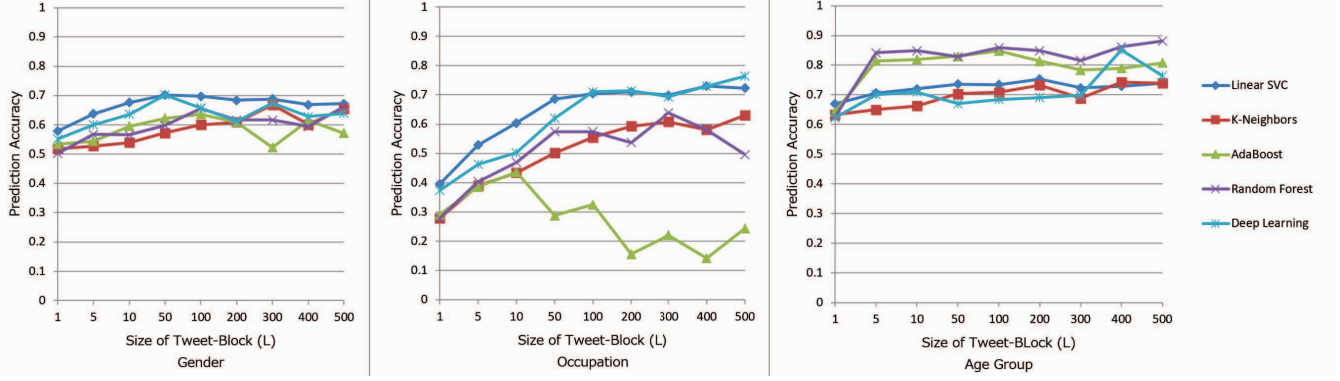


Figure 4. Prediction accuracy as a function of L under the condition where the best N is selected for each attribute

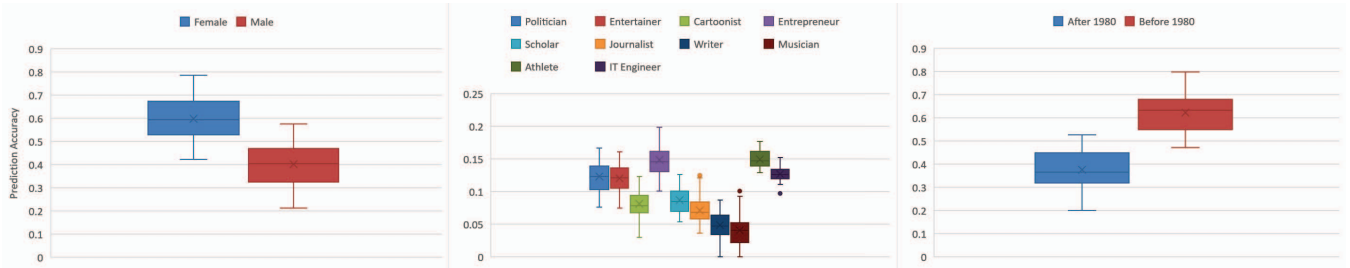


Figure 5. Predictability within personal attributes

group prediction, Linear SVC had a 75% accuracy at $N = 500$ and $L = 200$ (F-Score = 0.778 and AUC-Score = 0.749; digital immigrants (positive) / digital natives (negative)), and deep learning showed approximately about 80% accuracy at $N = 200$ and $L = 300$ (F-Score = 0.821 and AUC-Score = 0.839; digital immigrants (positive) / digital natives (negative)). In most cases, a larger value for L and N led to a little higher accuracy, but this soon reached a plateau.

Figure 4 shows the prediction accuracy as a function of L under the condition where the best N was selected for each attribute. For gender and occupation predictions, the best accuracy was achieved using Linear SVC and deep learning. In contrast, for age groups the best accuracy was achieved using Random Forest and AdaBoost. This result suggests that the optimal algorithm varies with the kinds of personal attributes, although Linear SVC and deep learning showed better and stable performance in our experiments.

B. Differences in Predictability within Personal Attributes

With regard to the three personal attributes, some entities have been thought to be more easily predictable than others. Intuitively, for example, politicians should be more predictable than others, since their word choice and usage would be unique. Figure 5 shows accuracy

distributions for each attribute entities computed from all combinations of N and L in deep learning. In terms of the inference of personal attributes from texts, females were easier to predict than males (t -test, $P < 0.001$), entrepreneurs were easier to predict than musicians (t -test, $P < 0.001$), and digital immigrants were easier to predict than digital natives (t -test, $P < 0.001$).

V. DISCUSSION

In this paper, we reported two preliminary results regarding the prediction of personal attributes (gender, jobs, and age groups) from text tweets by five different machine learning algorithms.

The results showed that the prediction of age groups is easier than the other two attributes. This can be explained by greater differences in word choice and usage between digital natives and digital immigrants, although this needs to be validated in the future. All the algorithms exceeded 60% accuracy for age group prediction, which was achieved even at $L = 1$ (i.e., a single tweet) if selecting for a proper value of N . As for gender and occupation predictions, larger values of L and N were required to obtain more than 60% accuracy ($N = 500$ and $L = 50$ for gender, $N = 500$ and $L = 100$ for occupation). Both predictions lead to approximately 70% accuracy. It is encouraging that 60–70% accuracy could be obtained for inferring personal attributes based

only on text posts. This suggests that much higher accuracy could be achieved by adding other features, such as images or URLs embedded in tweets.

In this study, deep learning did not show significant advantages compared with Linear SVC, but it has a potential advantage. While linear SVC needs to learn each attribute separately, deep neural networks (DNNs) can learn multiple attributes at the same time. If those attributes are mutually dependent, DNNs could learn faster with better prediction performance. This also needs to be tested in the future.

Compared to single tweets, tweet blocks significantly improved accuracy for all prediction tasks, suggesting that the prediction of personal attributes require a certain number of tweets. In other words, tweet blocks as relatively low dimensional vectors can convey enough personal information such that personal attributes can be algorithmically inferred. According to our study, $L = 50$ is the lower limit for gender and occupation predictions of approximately 60% accuracy. As shown in Figs. 2 and 4, even fewer tweets can yield a higher accuracy for the prediction of age groups.

Regarding the effects of the embedding dimension (N) for word2vec, better results were achieved in most cases when $N = 500$ rather than $N = 50$. In reality, however, greater N requires larger computational costs. Thus, it is better to select an appropriate N to balance computational costs and desired performance.

For future research, we will conduct a prediction test for other personal attributes by improving deep learning algorithms with a larger dataset. Once established a general framework for the inference of personal attributes can be applied to a wide variety of fields, including basic research on computational social science and applications for personalization and marketing.

ACKNOWLEDGMENT

This research was supported by JST PRESTO Grant Number JPMJPR16D6, JST CREST Grant Number JPMJCR17A4, and JSPS/MEXT KAKENHI Grant Numbers JP15H03446 and JP17H06383 in #4903.

REFERENCES

- [1] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2518–2525.
- [2] H. Lin, J. Jia, Q. Guo, Y. Xue, J. Huang, L. Cai, and L. Feng, "Psychological stress detection from cross-media microblog data using deep sparse neural network," in *Proceedings of 2014 IEEE International Conference on Multimedia and Expo (ICME)*, 2014, pp. 1–6.
- [3] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level psychological stress detection from social media using deep neural network," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 507–516.
- [4] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Alstynne, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.
- [5] S. A. Golder and M. W. Macy, "Digital footprints: Opportunities and challenges for online social research," *Annual Review of Sociology*, vol. 40, 2014.
- [6] <https://about.twitter.com/ja/company/>.
- [7] <https://twitter.com/TwitterJP/status/793649186935742465/>.
- [8] "Twitter API," <https://developer.twitter.com/en/docs>.
- [9] S. A. Golder, S. A. Golder, M. W. Macy, and M. Macy, "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science*, vol. 333, no. 6051, pp. 1878–1881, 2011.
- [10] K. Sasahara, Y. Hirata, M. Toyoda, M. Kitsuregawa, and K. Aihara, "Quantifying collective attention from tweet stream," *PLoS ONE*, vol. 8, no. 4, p. e61823, 2013.
- [11] Y. Takeichi, K. Sasahara, R. Suzuki, and T. Arita, "Concurrent bursty behavior of social sensors in sporting events," *PLoS ONE*, vol. 10, no. 12, pp. e0144646–13, Dec. 2015.
- [12] R. Kaur and K. Sasahara, "Quantifying moral foundations from various topics on twitter conversations," in *Proceeding of the 2016 IEEE Conference on Big Data*, 2016, pp. 2505–2512.
- [13] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The Development and Psychometric Properties of LIWC2007."
- [14] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.
- [15] L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana, "Knowing the tweeters: Deriving sociologically relevant demographics from twitter," *Sociological Research Online*, vol. 18, no. 3, p. 7, 2013.
- [16] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," *PLoS ONE*, vol. 10, p. e0115545, 2015.
- [17] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dzirzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, p. e73791, 2013.

- [18] M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.
- [19] Y. Wang, Y. Feng, Z. Hong, R. Berger, and J. Luo, “How polarized have we become? a multimodal classification of trump followers and clinton followers,” vol. 10539, 2017, pp. 440–456.
- [20] X. Liu and T. Zhu, “Deep learning for constructing microblog behavior representation to identify social media user’s personality,” *PeerJ Computer Science*, vol. 2, p. e81, 2016.
- [21] “IBM Watson Personality Insights,” <https://www.ibm.com/watson/services/personality-insights/>.
- [22] “Mecab: Yet another part-of-speech and morphological analyzer,” <http://mecab.sourceforge.net>, 2005.
- [23] “mecab-ipadic-neologd: Neologism dictionary for mecab,” <https://github.com/neologd/mecab-ipadic-neologd>.
- [24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv:1301.3781*, 2013.
- [25] “A powerful, flexible, and intuitive framework for neural networks,” <https://chainer.org>.
- [26] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1188–1196.
- [27] “scikit-learn: Machine learning in python,” <http://scikit-learn.org/stable/>.