Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 116 (2017) 604–611

Procedia Computer Science

www.elsevier.com/locate/procedia

2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017, Bali, Indonesia

Personality Prediction System from Facebook Users

Tommy Tandera, Hendro, Derwin Suhartono*, Rini Wongso, and Yen Lina Prasetio

Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K. H. Syahdan No. 9 Kemanggisan, Jakarta 11480, Indonesia

Abstract

The use of social networks is increasing rapidly. Various informations are shared widely through social media, i.e. Facebook. Information about users and what they expressed through status updates are such important assets for research in the field of behavioral learning and human personality. Similar researches have been conducted in this field and it grows continually till now. This study attempts to build a system that can predict a person's personality based on Facebook user information. Personality model used in this research is Big Five Model Personality. While other previous researches used older machine learning algorithm in building their models, this research tries to implement some deep learning architectures to see the comparison by doing comprehensive analysis method through the accuracy result. The results succeeded to outperform the accuracy of previous similar research with the average accuracy of 74.17%.

Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.

Keywords: personality prediction; big five model personality; facebook; machine learning; deep learning

1. Introduction

Social media has become the most widely used communication and interaction tool between people over the past few years. Direct interaction between people is decreasing as people tend to communicate indirectly through smartphones. Thus, it is quite difficult to recognize person's personality. However, what's written in social media might help us to get the information needed as people spend much time checking social media and expressing their feelings and thoughts through statuses, comments, and updates. Facebook has the largest users reaching 1.8 billion

* Corresponding author. Tel.: +62-21-5345830; fax: +62-21-5300244.
E-mail address: dsuhartono@binus.edu

1877-0509
Peer-review under responsibility of the scientific committee of the 2nd International Conference on Computer Science and Computational Intelligence 2017.
10.1016/j.procs.2017.10.016

users with around 800 million users spending about 40 minutes a day using it [1]. Facebook users generally express their feelings and opinions through status updates or comments. Although Facebook is currently more widely used to share photos and videos, this research focuses on users' linguistic aspect which is their status updates. Studies in the field of psychology showed that there is a correlation between personality and the linguistic behavior of a person [2,3]. This correlation can be effectively analyzed and illustrated using natural language processing approach. Therefore, the goal of this research is to build a prediction system that can automatically predict user personality based on their activities in Facebook.

There are several personality models used in predicting personality, such as Big Five Personality, MBTI (Myers-Briggs Type Indicator) or DISC (Dominance Influence Steadiness Conscientiousness). However, after some considerations and literature review process, Big Five Personality is used in this study as it is the most popular and precise in telling someone's personality traits. Traits in this model consist of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

The corpus used in this study consists of 2 (two) datasets. The first dataset consists of 250 users with around 10,000 statuses obtained from *myPersonality* project sample data, and the second dataset consists of 150 users which are collected manually. Prediction system is built using some linguistic features with different approach. The first used closed vocabulary that includes some features such as LIWC (Linguistic Inquiry and Word Count) and SPLICE (Structured Programming for Linguistic Cue Extraction). SNA (Social Network Analysis) is also included in the process because all the features' scores are provided by *myPersonality* dataset. All features in the first approach are specifically used in the older machine learning algorithm implementation. The second approach used open vocabulary approach. It is word embedding which is specifically used in deep learning technique implementation. This study uses some machine learning algorithms which are widely used in previous researches. To the best to our knowledge, deep learning implementation in this field is still hard to find. Hence, we implement some deep learning algorithms so that improvement to the prediction system can be made.

2. Related Work

Previous study on personality prediction has been done by using social media Facebook and some features such as LIWC features, SNA features, time-related features, and others [4]. Their research is very similar with ours especially for the dataset (250 dataset from myPersonality) and the features (LIWC and SNA features). Another research in personality prediction based on Facebook status were done by using two approaches such as open-vocabulary DLA (Differential Language Analysis) and LIWC features [5]. By using Facebook, a research defining features with bag-of-words and token (unigrams) approaches were conducted as well. Other study was done to make a personality prediction system by using Twitter with LIWC and MRC as features [6].

All mentioned above researches did personality prediction by using social media in English based on Big Five Personality models. Recent research was conducted to make a personality prediction system using Twitter in Bahasa based on Big Five Personality models [7,8]. Other research on personality prediction was done using deep learning technique to classify Big Five Personality models from social media Facebook [9].

3. Methodology

3.1. Dataset

The dataset used in this study is divided into two parts. The first dataset obtained from myPersonality [10] consists of 250 data of Facebook users with approximately 10,000 statuses with given personality label based on the Big Five Personality Traits model. The distribution of the myPersonality dataset based on the personality type is presented in Table 1 below.

Table 1. Distribution of myPersonality dataset.

Value	OPN	CON	EXT	AGR	NEU
Yes	176	130	96	134	99
No	74	120	154	116	151

The second dataset is the statuses of 150 Facebook users which are collected manually. Facebook API Graph is utilized in the process of collecting the dataset. Personality labeling is then done by manually entering the user posts into Apply Magic Sauce application (https://applymagicsauce.com/). Apply Magic Sauce application is a web application build by Cambridge Psychometrics Centre to predict psychological traits from digital footprints of human behavior. Their models are based on over 6 million social media profiles and matching scores on psychometrics tests. They have published their methods in the Proceedings of the National Academy of Sciences [11] and proven to predict someone better than their friends or partners [12]. Apply Magic Sauce open for any researches that want to use their API to help them collect information on psychological characteristics based on Big Five Personality without inconveniencing the participants with personality questionnaires. Table 2 is the result of dataset distribution after being labeled based on Big Five Personality Traits model using Apply Magic Sauce.

Table 2. Distribution of Manual gathered dataset.

Value	OPN	CON	EXT	AGR	NEU
Yes	97	63	38	81	50
No	53	87	112	69	100

3.2. Features

This study uses several features to see the comparison of the results and capabilities between them. The main reason is to investigate the suitability and performance of this various features for personality modeling. The features used are differentiated for each learning implementation. For traditional machine learning implementation, we used linguistic feature with closed-vocabulary approach. Closed vocabulary is a feature based on the number of words content in accordance with predefined features. For this approach, we used linguistic features such as LIWC [13] and SPLICE [14]. LIWC used in this study was LIWC2015 version which had 85 features enhanced from LIWC2007. All LIWC features were used. SPLICE is a linguistic feature that has been used in several studies in this field [14]. 74 features of SPLICE are used in this research. In addition to the linguistic features described beforehand, this research also utilized the use of SNA features provided by myPersonality dataset in form of detail information about a user's friendship network [15].

In contrast to the implementation of traditional machine learning, deep learning utilization was done separately by using linguistic features of open vocabulary approach. Open vocabulary does not require predefined features. This approach performs an automatic exploration of dataset to find relationship between words with personality. The actual technique used in this study is word embedding using Glove [16] which has around 6 billion tokens, 400 thousand words, and 100 vector dimensions. Previous studies comparing these 2 (two) linguistic features approaches have been done before [5].

3.3. Preprocessing

All data in form of English went through the preprocessing stage before it could be processed. Pre-preprocessing steps consist of removing URLs, symbols, names, spaces, lowering case, stemming, and removing stop words. Whilst, data in Bahasa went through additional preprocessing process; it was replacement of slang words or non-standard words which was manually conducted. After finished, it was then translated into English.

Steps such as removing names, stop words and stemming were using NLTK library. 153 stop words were removed in the experiments. Another process was done manually by using written regex and codes.

3.4. Model classification

As mentioned above, traditional machine learning and deep learning were used in classification process. Traditional machine learning algorithms included Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis (LDA). In this research, we consider all machine learning algorithms used above as the traditional machine learning as it will be differentiated and compared to deep learning algorithm which is actually subset from machine learning methods but it differs in the use of neural networks that contain more than one hidden layer. For model validation, researchers used a 10-fold cross validation technique using Python libraries. 10-fold cross validation divided 10% dataset into data testing and 90% dataset as training data in turn.

We conducted a series of tests with various scenarios to see the accuracy of each algorithm in predicting personality type. Testing was done by adding some additional processes to improve accuracy. The first process was Features Selection that try to filter or remove the features that were considered to have a low correlation to the traits of the personality. The correlation value was calculated using chi-square method. We did some "try and error" experiments until we found the best setting for this process. The next process was resampling that aimed to balance the data because of the unbalanced data distribution. As shown in Table 1, Openness traits have a comparison of binary classes 2.4 (yes): 1 (no). In Table 2, it is found that Extraversion traits have a binary class comparison of 1 (yes): 2.9 (no). The resampling technique was Under-sampling and Over-sampling. These techniques were applied using library from imbalanced_learn that include SMOTE function for Over-sampling as well as ClusterCentroids function for Under-sampling.

Meanwhile, deep learning implementations were using four architectures, namely MLP (Multi-Layer Perceptron), LSTM (Long Short Term Memory), GRU (Gated Recurrent Unit), and CNN 1D (1-Dimensional Convolutional Neural Network). We attempted to combine LSTM and CNN 1D architecture as an additional architecture as well. MLP consists of input, hidden, and output layers which is using a basic algorithm for training, known as backpropagation [17]. CNN 1D has the same layers as MLP, but there are two more layers before the MLP's layers namely convolutional layers and max pooling layers [18]. GRU is a simplification of LSTM, they used peephole connections and output activation functions, and coupled the input and the forget gate into an update gate [19]. A series of scenarios were conducted to obtain the highest prediction's accuracy for each architecture. The testing was done by adding the resampling process. The Python library used is Keras and Theano as the backend. However, for this implementation we did not do the validation using 10-fold cross in the testing process yet due to the limitation of hardware capability which caused out of memory problem. So, we figured out the solution using parting the dataset into training dataset and testing dataset with the distribution of 80% and 20% from the total data. This allocation of testing data was randomly selected and we obtained 50 datasets as testing data from myPersonality and 30 datasets as testing data from manual gathered dataset.

Table 3 is a breakdown of experimental scenarios to be performed on traditional machine learning and deep learning.

Table 3. Experimental scenarios for traditional machine learning and deep learning.

	Machine Learning							
	Features			Feature Selection		Resampling		
Scenario	LIWC	SPLICE	SNA	No	Yes	Without Resampling	Under-sampling	Over-sampling
1	✓			✓		✓		
2	✓			✓			✓	
3	✓			✓				✓
4	✓				✓	✓		
5	✓				✓		✓	
6	✓				✓			✓
7		✓		✓		✓		
8		✓		✓			✓	
9		✓		✓				✓

10		✓			✓	✓		
11		✓			✓		✓	
12		✓			✓			✓
13			✓	✓		✓		
14			✓	✓			✓	
15			✓	✓				✓
16			✓		✓	✓		
17			✓		✓		✓	
18			✓		✓			✓

	Deep Learning		
	Resampling		
Scenario	Without Resampling	Under-sampling	Over-sampling
19	✓		
20		✓	
21			✓

4. Classification Result

All clasification results obtained by using traditional machine learning and deep learning can be seen in Table 4, 5, 6, and 7. We only report the algorithms, architectures, and scenario number with the highest accuracy in each traits.

Table 4 shows the result obtained by using myPersonality dataset. The implementation of traditional machine learning which shows the highest accuracy is dominated by scenario number 1 and 4. The highest accuracy is 70.40% obtained by using SVM and Logistic Regression algorithm. The highest average accuracy is 63.04% obtained by using LDA algorithm. The highest average accuracy for all traits is 68.80% obtained from Openness (OPN).

Table 5 shows the result obtained by using Manual gathered dataset and implementation machine learning shows the highest accuracy is dominated by scenario number 1 and 4. The highest accuracy is 79.33% obtained by using LDA algorithm. The highest average accuracy is 67.20% obtained by using SVM algorithm. The highest average accuracy for all traits is 75.87% obtained from Extraversion (EXT).

Meanwhile, Table 6 shows the result obtained by using myPersonality dataset. Implementation of deep learning shows the highest accuracy is dominated by scenario number 20. The highest accuracy is 79.49% obtained by using MLP architecture. The highest average accuracy is 70.78% obtained by using MLP architecture. The highest average accuracy for all traits is 74.10% obtained from Openness (OPN).

Table 7 shows the result obtained by using manual gathered dataset. Implementation of deep learning shows the highest accuracy is dominated by scenario number 21. The highest accuracy is 93.33% obtained by using MLP and LSTM+CNN 1D architecture. The highest average accuracy is 74.17% obtained by using LSTM+CNN 1D architecture. The highest average accuracy for all traits is 83.33% obtained from Extraversion (EXT).

By observing all average accuracies from experiments on traditional machine learning, it is found that the accuracy is quite balanced. However in the deep learning implementation, all average accuracies for each architecture is quite different. From the average accuracy result based on traits, we can see Openness (OPN) has the highest average accuracy in myPersonality dataset, while Extraversion (EXT) has the highest average accuracy in manual gathered dataset. This result may be different in other research, as it heavily depends on the dataset and the classification model with its features that being used for the prediction. Based on the experimental results, we can conclude that the highest average accuracy is obtained by using implementation deep learning but there is no architecture that dominated all big 5 personality traits.

Table 4. Traditional machine learning classification result by using myPersonality dataset *

Algorithm	Traits (Scenarios)					Average
	OPN	CON	EXT	AGR	NEU	
Naive Bayes	70.00% (4)	59.20% (14)	68.80% (1)	56.40% (8)	54.40% (1)	61.76%
SVM	70.40% (4)	56.00% (4)	61.60% (4)	56.80% (12)	60.40% (4)	61.04%
Logistic Regression	70.40% (1)	54.40% (3)	68.40% (1)	53.60% (5)	60.40% (4)	61.44%
Gradient Boosting	63.20% (1)	56.40% (5)	68.00% (13)	63.20% (6)	59.20% (16)	62%
LDA	70.00% (16)	58.40% (14)	68.00% (16)	58.00% (7)	60.80% (1)	63.04%
Average	68.80%	56.88%	66.96%	57.60%	59.04%	

Table 5. Traditional machine learning classification result by using manual gathered dataset *

Algorithm	Traits (Scenarios)					Average
	OPN	CON	EXT	AGR	NEU	
Naive Bayes	60.67% (1)	62.67% (1)	73.33% (1)	53.33% (2)	70.00% (4)	64.00%
SVM	64.67% (4)	65.33% (1)	76.00% (1)	60.67% (12)	69.33% (1)	67.20%
Logistic Regression	65.33% (7)	66.67% (11)	74.67% (4)	59.33% (5)	66.67% (1)	66.53%
Gradient Boosting	67.33% (1)	62.67% (1)	76.00% (4)	58.67% (7)	66.67% (1)	66.26%
LDA	60.00% (4)	67.33% (1)	79.33% (1)	60.67% (3)	66.67% (4)	66.80%
Average	63.60%	64.93%	75.87%	58.53%	67.87%	

Table 6. Deep learning classification result by using myPersonality dataset *

Architectures	Traits (Scenarios)					Average
	OPN	CON	EXT	AGR	NEU	
MLP	79.31% (20)	59.62% (21)	78.95% (20)	56.52% (20)	79.49% (20)	70.78%
LSTM	68.00% (19)	52.00% (19)	58.00% (19)	56.52% (20)	58.62% (21)	58.63%
GRU	68.00% (19)	62.00% (19)	58.00% (19)	65.22% (20)	64.00% (19)	63.44%
CNN 1D	79.31% (20)	50.00% (21)	60.94% (21)	67.39% (20)	61.54% (20)	63.84%
LSTM+CNN 1D	75.86% (20)	57.69% (21)	71.05% (20)	50.00% (21)	58.97% (20)	62.71%
Average	74.10%	56.26%	65.39%	59.13%	64.52%	

Table 7. Deep learning classification result by using Manual gathered dataset *

Architectures	Traits (Scenarios)					Average
	OPN	CON	EXT	AGR	NEU	
MLP	66.67% (20)	64.00% (20)	93.33% (20)	70.37% (20)	75.00% (20)	73.87%
LSTM	67.50% (21)	64.00% (20)	70.00% (19)	66.67% (20)	75.00% (20)	68.63%
GRU	63.33% (19)	61.76% (21)	73.33% (20)	59.38% (21)	76.67% (19)	66.89%
CNN 1D	76.19% (20)	68.00% (20)	86.67% (20)	63.33% (19)	75.00% (20)	73.84%
LSTM+CNN 1D	67.50% (21)	66.67% (19)	93.33% (20)	63.33% (19)	80.00% (20)	74.17%
Average	68.24%	64.89%	83.33%	64.62%	76.33%	

* Number in brackets for each trait indicates the number of scenario in Table 3.

In this research, we did experiment on personality prediction based on Big Five Personality models using traditional machine learning and deep learning to classify the traits. For traditional machine learning, we used five algorithms. They are Naive Bayes, SVM, Logistic Regression, Gradient Boosting, and LDA with three features which are LIWC, SPLICE, and SNA. 10-fold cross validation is used for the evaluation model. The experimental

scenarios are done by using 2 datasets, feature selection, and resampling. Experimental scenario by using myPersonality dataset shows that the highest accuracy is 70.40% obtained by using SVM and Logistic Regression algorithm for Openness (OPN) trait with LIWC features. SVM algorithm was used with feature selection while Logistic Regression was used without feature selection. Both of them were used with no resampling. Experimental scenario by using manual gathered dataset shows that the highest accuracy is 79.33% obtained by using LDA algorithm for Extraversion (EXT) trait with LIWC features, without feature selection, and without resampling.

The results of experiments on traditional machine learning proved that LDA algorithm has the highest average accuracy in myPersonality dataset and SVM algorithm has the highest average accuracy in manual gathered dataset but not much different from other algorithms. LIWC without feature selection has the highest accuracy among other features in both dataset. We also performed a combination of LIWC, SPLICE, and SNA yet it still can not improve the accuracy. Resampling technique can not improve the accuracy as well.

For deep learning implementation, we used 4 architectures; they are MLP, LSTM, GRU, and CNN 1D. We also tried to combine LSTM with CNN 1D architecture. Experimental scenarios are done using two datasets and resampling. Experimental scenario by using myPersonality dataset shows that the highest accuracy is 79.49% obtained by using MLP architecture for Openness (OPN) trait with resampling (under-sampling technique). Experimental scenario by using manual gathered dataset shows that the highest accuracy is 93.33% by using MLP and LSTM+CNN 1D architectures for Extraversion (EXT) trait with resampling (under-sampling technique).

The results of experiments on deep learning proved that MLP architecture has the highest average accuracy in myPersonality dataset and LSTM+CNN 1D architectures has the highest accuracy in manual gathered dataset. In addition, resampling technique can also improve the accuracy significantly especially under-sampling technique.

5. Conclusion

The results of experiments show that deep learning can improve the accuracy even if the accuracy is still quite low for some traits. It is possibly due to small number of dataset used in this study. However, the results of this study by using traditional machine learning and deep learning can outperform the results of previous studies using the same dataset.

Hence, for future study, we plan to collect and build more dataset. We also plan to use XGBoost algorithm [20], other architectures, and other processes to improve this prediction system.

References

1. Bachrach Y, Kosinski M, Graepel T, Kohli P, Stillwell D. Personality and patterns of Facebook usage. In 4th Annual ACM Web Science Conference; 2012. p. 24-32.
2. Mairesse F, Walker M, Mehl M, Moore R. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research. 2007; 30: p. 457-500.
3. Fast L, Funder D. Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. Journal of personality and social psychology. 2008; 94(2): p. 334.
4. Farnadi G, Zoghbi S, Moens M, De Cock M. How well do your Facebook status updates express your personality? In 22nd edition of the annual Belgian-Dutch conference on machine learning (BENELEARN); 2013. p. 88.
5. Schwartz H, Eichstaedt J, Kern M, Dziurzynski L, Ramones S, Agrawal M, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one. 2013; 8(9).
6. Wijaya A, Febrianto N, Prasetia I, Suhartono D. Sistem Prediksi Kepribadian "The Big Five Traits" Dari Data Twitter. Jakarta: Bina Nusantara University, School of Computer Science; 2016.
7. Ong V, Rahmanto ADS, W, Suhartono D. Exploring Personality Prediction from Text on Social Media: A Literature Review. Internetworking Indonesia Journal. 2017; 9(1): p. 65-70.
8. Ong V, Rahmanto ADS, W, Suhartono D, Nugroho AE, Andangsari EW, et al. Personality Prediction Based on Twitter Information in Bahasa. In 2nd International Workshop on Language Technologies and Applications (LTA'17); 2017; Prague.
9. Majumder N, Poria S, Gelbukh A, Cambria E. Deep Learning-Based Document Modeling for Personality Detection from Text. IEEE Intelligent Systems. 2017 Mar; 32(2)(IEEE): p. 74-79.
10. Kosinski M, Matz S, Gosling S, Popov V, Stillwell D. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American Psychologist. 2015 Feb; 70(6): p. 543.
11. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. In Proceedings of the National Academy of Sciences of the United States of America; 2013: PNAS. p. 5802-5805.

12. Youyou W, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans. In National Academy of Sciences; 2015. p. 1036-1040.
13. Pennebaker J, Boyd R, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. 2015..
14. Moffitt K, Giboney J, Ehrhardt E, Burgoon J, Nunamaker J. Structured programming for linguistic cue extraction. [Online].; 2010. Available from: http://splice.cmi.arizona.edu/.
15. O'malley A, Marsden P. The analysis of social networks. In Health services and outcomes research methodology; 2008. p. 222-69.
16. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In EMNLP 2014; 2014. p. 1532-1543.
17. Gardner M, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment. 1998; 32(14): p. 2627-2636.
18. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. 2014..
19. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014..
20. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785-794.