Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2017) 000–000

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

Big Five Personality Prediction System from Facebook Users

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

Bina Nusantara University, Jl. K.H. Syahdan No. 9, Jakarta 11480, Indonesia

Abstract

The usage of social networks has now reached its peak. Various information shared widely through social media such as Facebook. Information about users and their statuses is such an important asset for research in the field of behavioral learning and human personality. Similar researches have been conducted in this field and continue to grow to date. This study attempts to build a system that can predict a person's personality based on Facebook user information. Personality model that used in this research is Big Five Model Personality. While other previous research using older machine learning algorithm in building their model, this research try to implement some deep learning architectures to see the comparison by doing comprehensive analysis method through the accuracy result. The results shown in this study succeeded to outperform the accuracy of previous similar research with the current highest accuracy of 93.33% acquire using deep learning architecture.

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

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

1. Introduction

Social media has become the most used communication and interaction tool between people over the past few years. In the era where almost all human beings have their own smartphones, direct interaction between people almost rarely happens. So, it is quite difficult to recognize and get to know the personality of a person. However, this is totally different from what happens in social media. Facebook has the largest users reaching 1.8 billion users with aroung 800 million users spending about 40 minutes a day using Facebook [1]. Facebook users generally express their feelings and opinions in their user feed. Although Facebook is currently more widely used to share photos and

* Corresponding author. Tel.: +62-21-5345830; fax: +62-21-5300244.
E-mail address: tommy.tandera@binus.ac.id; hendro004@binus.ac.id; dsuhartono@binus.edu; rwongso@binus.edu; yenlina@binus.edu

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

videos, this research will be focus on users' linguistic aspect which is their statuses. Various studies in the field of psychology show that there is a correlation between personality and the linguistic behavior of a person. This correlation can be effectively analyzed and illustrated using NLP approach. Therefore this research goal is to build a prediction system that can automatically predict an user personality based on their activity in Facebook.

This prediction system will be built using the Big Five Personality model. There are several other personality models used in related study such as MBTI (Myers-Briggs Type Indicator) or DISC. However, after some considerations and literature review process, Big Five Personality is chosen by the reason it's 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 in this study will be divided by two datasets. First dataset consists of 250 users with around 10,000 statuses obtained from myPersonality project sample data and the other dataset of 150 users collected manually. Prediction system will be built using some linguistic features with different approach. The first is using closed vocabulary that includes some features such as LIWC and SPLICE. SNA (Social Network Analysis) also included in the process because all the features score provided by myPersonality dataset. All features in the first approach specifically used in the older machine learning algorithm implementation. The second approach is using open vocabulary approach which will be word embedding features specifically used in deep learning technique implementation. Similar research using machine learning older algorithm which is we'll also use in this research has been widely used before, but the implementation of deep learning in this field of research still hardly to find. Therefore, this research will also conduct the implementation of deep learning to see whether it can boost the result of the prediction system. The best classifier and features will be based on the accuracy result and will be used as the model for the final personality prediction system.

2. Related Work

Previous study on personality prediction has been done by [2] using social media Facebook and some features such as LIWC features, SNA features, time-related features, and others. [2] is very similar with our study especially the dataset (250 dataset from myPersonality) and the features (LIWC and SNA features). [3] has researched on personality prediction based on Facebook status by using two approaches such as open-vocabulary DLA (Differential Language Analysis) and LIWC features. [4] also has researched by using Facebook with bag-of-words and token (unigrams) approaches as features. Other study has been done by [5] to make a personality prediction system by using Twitter with LIWC and MRC as features. [2], [3, [4], dan [5] have researched on personality prediction by using social media in english based on Big Five Personality models. Recent research has been done by [6] to make a personality prediction system using Twitter in bahasa based on Big Five Personality models. Other research on personality prediction also has been done by [7] using deep learning technique to classify Big Five Personalitiy models from social media Facebook.

3. Methodology

3.1. Dataset

The dataset used in this study is divided into two parts. The first dataset obtained from myPersonality [8] as many as 250 datasets of Facebook users with approximately 10,000 statuses that have been given labeling personality 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 status of 150 Facebook user datasets 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 applymagicsauce app. Table 2 is the result of dataset distribution after being labeled based on Big Five Personality Traits model.

Table 2. Distribution of Manual Gathering dataset.

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

3.2. Features used

This study will use several features to see the comparison of 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 of the learning implementations. For machine learning implementation, we use 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 [9] and SPLICE [10]. LIWC used in this study is LIWC2015 version which has 85 features that have been developed from LIWC2007 version. In this study all LIWC features will be used.

SPLICE is a linguistic feature created by Moffit et al and has been used in several studies in this field. In this study there are 74 features of SPLICE that will be used.

In addition to the above linguistic features, this research will also utilize the use of Social Network Analysis features provided by the myPersonality dataset in the form of detailed information about a user's friendship network. For complete information on this feature can be seen in [11].

In contrast to the implementation of machine learning, implementation of deep learning was done separately by using linguistic features of open vocabulary approach. Open vocabulary does not require predefined features. This approach will perform an automatic exploration of the dataset used to find the relationship between the uses of words with personality. The actual technique that used in this study is word embedding using GloVe. GloVe that we are using have around 6 billion tokens, 400 thousand words, and 100 vector dimensions. Previous studies that have made comparisons between these two linguistic feature approaches have been done in [3].

3.3. Preprocessing

All data that has been collected in this research will go through the preprocessing stage before build the classification model. Pre-preprocessing steps are removing URLs, remove symbols, remove names, remove spaces, lower case text, stemming, and remove stopwords. Especially for status with Indonesian language, additional preprocessing process is done manually by replacing slang words or non-standard words from the status first to then proceed to the translation into English.

Steps such as removing names, stopwords and stemming using NLTK library in the process. There are 153 stopwords that removed in this research. The other steps are done manually using written regex and codes.

3.4. Model classification

Implementation of machine learning using 5 different algorithms, namely Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and Linear Discriminant Analysis (LDA). For model validation, researchers used a 10-fold cross validation technique using Python libraries. 10-fold cross validation divides 10% dataset into data testing and 90% dataset as training data in turn.

Researchers conducted a series of tests with various scenarios to see how the algorithm accuracy results in predicting the personality type. Testing is done by adding some additional processes to improve accuracy. The first process is Features Selection that tries to filter or remove the features used that are considered to have a low correlation to the traits of the personality. The correlation value is calculated using chi-square method. We did try and error experiment until we find the best setting for this process. The next process is to do a resampling process that aims to balance the distribution of data where the data distribution on the personality type has an unbalanced distribution as in Table 1 where Openness traits have a comparison of binary classes 2.4 (yes): 1 (no) and Table 2 where Traits Extraversion has a binary class comparison of 1 (yes): 2.9 (no). The resampling technique used is Under-sampling and Over-sampling. These techniques applied using library from imbalanced_learn that include SMOTE function for Over-sampling as well as ClusterCentroids function for Under-sampling.

Implementation of deep learning using four architectures, namely MLP, LSTM, GRU, and CNN 1D. Then the researchers tried to combine LSTM and CNN 1D architecture as an additional architecture. Researchers conducted a series of scenarios to obtain the highest prediction accuracy for each architecture. The test is done by adding the resampling process. The Python library used is Keras and Theano as the backend. However, for this implementation we haven’t done the validation using 10-fold cross in the testing process due to lack of hardware capability that causes out of memory problem. So, we figured out the solution using parting the dataset into training dataset and testing dataset with the distribution quantity of 80%-20% from the total data. This allocation of testing data is randomly selected that took out 50 dataset as testing data from myPersonality dataset and 30 dataset as testing data from manual gathering dataset.

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

Table 3. Experimental scenarios for 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 result by using machine learning and deep learning is shown in Table 4, 5, 6, and 7. We only show the algorithms, architectures, and scenario number with the highest accuracy in each traits.

Table 4 by using myPersonality dataset and implementation machine learning shows the highest accuracy is dominated by scenario 1 and 4. The highest accuracy is 70.40% obtained by using SVM algorithm 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 by using Manual Gathering dataset and implementation machine learning shows the highest accuracy is dominated by scenario 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).

Table 6 by using myPersonality dataset and implementation deep learning shows the highest accuracy is dominated by scenario 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 by using Manual Gathering dataset and implementation deep learning shows the highest accuracy is dominated by scenario 21. The highest accuracy is 93.33% obtained by using MLP architecture and LSTM+CNN 1D architectures. The highest average accuracy is 74.17% obtained by using LSTM+CNN 1D architectures. The highest average accuracy for all traits is 83.33% obtained from Extraversion (EXT).

For all average accuracy in each machine learning’s algorithms and each datasets show that the accuracy is quiet balanced. However in implementation deep learning, all average accuracy for each architectures is quiet different. While Extraversion (EXT) has the higher average accuracy than other traits. From 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. Machine learning classification result by using myPersonality dataset.
The number in the brackets for each trait indicates the scenario number in Table 3.

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. Machine learning classification result by using Manual Gathering dataset.
The number in the brackets for each trait indicates the scenario number in Table 3.

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.
The number in the brackets for each trait indicates the scenario number in Table 3.

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 Gathering dataset.
The number in the brackets for each trait indicates the scenario number in Table 3.

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%	

In this research, We have experimented on personality prediction based on Big Five Personality models. We implemented machine learning and deep learning to classify the traits.

In implementation machine learning, we used 5 algorithms which are Naive Bayes, SVM, Logistic Regression, Gradient Boosting, and LDA with 3 features which are LIWC, SPLICE, and SNA. 10-fold cross validation is used for the evaluation model. The experimental scenarios consist of 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 algorithm and Logistic Regression algorithm for Openness (OPN) trait with LIWC features. SVM algorithm with feature selection and Logistic Regression without feature selection and both algorithms without resampling. Experimental scenario by using Manual Gathering 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 machine learning prove that LDA algorithm has the highest average accuracy in myPersonality dataset and SVM algorithm has the highest average accuracy in Manual Gathering 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 but it can not improve the accuracy. Resampling technique also can not improve the accuracy.

In implementation deep learning, we used 4 architectures which are MLP, LSTM, GRU, and CNN 1D. We also tried to combine LSTM architecture with CNN 1D architecture. Experimental scenarios consist of using 2 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 Gathering dataset shows that the highest accuracy is 93.33% by using MLP architecture and LSTM+CNN 1D architectures for Extraversion (EXT) trait with resampling (under-sampling technique).

The results of experiments on deep learning prove that MLP architecture has the highest average accuracy in myPersonality dataset and LSTM+CNN 1D architectures has the highest accuracy in Manual Gathering dataset. Resampling technique also can improve the accuracy significantly especially under-sampling technique.

5. Conclusion

The results of experiments show that implementation deep learning can improve the accuracy. However, the accuracy is quiet low for some traits. We think because the number of dataset used in this study is still too small. But, the results of this study by using machine learning and deep learning can outperform the results of previous studies using the same dataset.

Hence, for future study, we plan to get more dataset from myPersonality. We also plan to use XGBoost algorithm, other architectures, and other processes to improve this prediction system.

References

[1] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell (2012) *Personality and Patterns of Facebook Usage*. ACM Web Science Conference. *Proceedings of the ACM Web Science Conference*, 36–44.

[2] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock (2013) *How Well Do Your Facebook Status Updates Express Your Personality?*. Conference on Machine Learning, Nijmegen, The Netherlands.

[3] H. Andrew Schwartz , Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, . . . Lyle H. Ungar (2013) *Personality, Gender, and Age in the Language of Social Media: The Open Vocabulary Approach*. PLOS ONE, 8, e73791.

[4] Firoj Alam, Evgeny A. Stepanov, and Giuseppe Riccardi (2013) *Personality Traits Recognition on Social Network – Facebook*. WCPR (ICWSM-13), Cambridge, MA, USA.

[5] Albert Wijaya, Nathanael Febrianto, Irwan Prasetia, and Derwin Suhartono (2016) *Sistem Prediksi Kepribadian "The Big Five Traits" Dari Data Twitter*. Bina Nusantara University, Jakarta, Indonesia.

[6] Veronica Ong, Anneke D. S. Rahmanto, Williem, and Derwin Suhartono (2017). *Personality Prediction Based on Twitter Information in Bahasa*. Internetworking Indonesia Journal, 9(1), 65-70.

[7] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, Erik Cambria (2017) *Deep learning-Based Document Modeling for Personality Detection from Text*. IEEE Intelligent Systems, 32(2), 74-79.

[8] *Michal Kosinski, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell (2015) Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. American Psychologist.*

[9] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn (2015). *The development and Psychometric Properties of LIWC2015*. University of Texas, Austin, Texas.

[10] Kevin C. Moffit, Justin S. Giboney, E. Ehrhardt, Judee K. Burgoon, and Jay F. Nunamaker (2012) *Structured Programming for Linguistic Cue Extraction (SPLICE)*. Report of the HICSS-45

[11] A. James O'Malley, Peter V. Marsden (2008) *The analysis of social networks*. Health Services and Outcomes Research Methodology, (8), 222–269.