Journal of Big Data

RESEARCH Open Access

Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging

Hans Christian[1], Derwin Suhartono[2*], Andry Chowanda[2] and Kamal Z. Zamli[3]

*Correspondence:
dsuhartono@binus.edu
[2] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
Full list of author information is available at the end of the article

Abstract

The ever-increasing social media users has dramatically contributed to significant growth as far as the volume of online information is concerned. Often, the contents that these users put in social media can give valuable insights on their personalities (e.g., in terms of predicting job satisfaction, specific preferences, as well as the success of professional and romantic relationship) and getting it without the hassle of taking formal personality test. Termed personality prediction, the process involves extracting the digital content into features and mapping it according to a personality model. Owing to its simplicity and proven capability, a well-known personality model, called the big five personality traits, has often been adopted in the literature as the de facto standard for personality assessment. To date, there are many algorithms that can be used to extract embedded contextualized word from textual data for personality prediction system; some of them are based on ensembled model and deep learning. Although useful, existing algorithms such as RNN and LSTM suffers from the following limitations. Firstly, these algorithms take a long time to train the model owing to its sequential inputs. Secondly, these algorithms also lack the ability to capture the true (semantic) meaning of words; therefore, the context is slightly lost. To address these aforementioned limitations, this paper introduces a new prediction using multi model deep learning architecture combined with multiple pre-trained language model such as BERT, RoBERTa, and XLNet as features extraction method on social media data sources. Finally, the system takes the decision based on model averaging to make prediction. Unlike earlier work which adopts a single social media data with open and close vocabulary extraction method, the proposed work uses multiple social media data sources namely Facebook and Twitter and produce a predictive model for each trait using bidirectional context feature combine with extraction method. Our experience with the proposed work has been encouraging as it has outperformed similar existing works in the literature. More precisely, our results achieve a maximum accuracy of 86.2% and 0.912 f1 measure score on the Facebook dataset; 88.5% accuracy and 0.882 f1 measure score on the Twitter dataset.

Keywords: Personality prediction, Natural language processing, Social media, Deep learning, BERT, Language model

Introduction

In recent years, information growth has proliferated in accelerating pace in line with the advent of social media especially in the form of textual data types. According to the Social Media Trend report published in [39], there are 3.8 billion active users of social media in the world as of January 2020, with a projected increase of 9.2% of users each year. Often, people use social media to express themselves on certain issues related to their lives and family well beings, psychology, financial issues, interaction with societies and environment, as well as politics. In some cases, these expressions can be used to characterize the individual behavior and personality. In fact, earlier studies (e.g. [4, 11, 16, 18, 24, 25]) demonstrate that there is a strong correlation between user personalities and their online behavior on social media. Some examples of applications that can take advantage from the user personality information include recruitment systems, personal counseling systems, online marketing, personal recommendation systems, and bank credit scoring systems to name a few [5, 12].

Owing to the inherent ambiguities of natural languages, developing an effective personality prediction model based on the textual message that user shares on social media can be a painstakingly difficult task. Dealing with these ambiguities, much progress has been achieved in the field of Natural Language Processing (NLP). To-date, NLP has enabled computers to understand words or sentences written in human language [6]. Linguistic concepts such as part-of-speech (nouns, verbs, adjectives) and grammatical structures are usually used in NLP [35]. Apart from part-of-speech and structural grammar, NLP is also able to deal with the anaphors and ambiguities that often arise in a language via knowledge representations such as a dictionary of words and their meanings, sentence structure, grammar rules, and other information such as synonyms or abbreviations [24].

Automatic personality prediction has become a widely discussed topic for researchers in the NLP community. References [13, 29] shows that personality can be defined as a pattern of influence or personality used to characterize unique individuals. The existing personality prediction exploited deep learning and machine learning algorithm along with open vocabulary feature extraction to improve classification accuracy. However, this approach has limitations to extract contextual features in the sentence due to limitedness of computational algorithm and out of vocabulary problem by using pre-defined corpus. Moreover, the small number dataset used in building personality prediction system especially using deep learning algorithm to be the main obstacle to maximize the model performance [5, 29, 37]. Addressing the aforementioned issues, this paper proposes a multi model deep learning architecture which build on top of different pre-trained language model called Bidirectional Encoder from Transformer (BERT), A Robustly Optimized BERT Pretraining Approach (RoBERTa), and XLNet a Generalized Autoregressive Pretraining for Language Understanding to capture the contextual meaning of a text-based data from social media. Later, the text-based data will be added with another additional NLP Features such as Sentiment Analysis, Term Frequency-Inverse Gravity Moment (TF-IGM), and National Research Council (NRC) Emotion Lexicon Database as features to build a multi model deep learning architecture to predict the personality traits. The contributions of this work can be summarized as follows:

- We proposed a multi model deep learning architecture with a pre-trained language model BERT, RoBERTa, and XLNet; along with additional NLP Features (sentiment analysis, TF-IGM, NRC emotion lexicon database) as features extraction method for personality prediction system.
- Unlike the other approach, we also proposed combining multiple sources of social media data to increase the number of datasets for better classification.
- We evaluate the performance of the model built and compare it with other previous studies algorithm that give the best performance in predicting personality.
- We show that our methods enable to produce better performance compare to the previous study in predicting personality traits.

Related works

Personality prediction using Facebook and Twitter dataset is not new. For example, research conducted by [18, 37, 38, 40] used an open-source Facebook personality dataset called MyPersonality which consists of 250 users with their status data and traits, and maps to big five personality model. Prevalent feature extraction method called Linguistic Inquiry and Word Count (LIWC), which is a linguistic analytical tool that helps in analyzing quantitative texts and provides a calculation number of words that have the meaning of categories based on a psychological dictionary is used as the main feature extraction method. The use of these analytical tools is increasingly popular as can be seen from the use of these methods in line with research conducted in the last two years. In addition, Social Network Analysis (SNA) is a technique in analyzing social structures that arise from a combination of people in a specific population and the interactions that occur Fsoftin that population. These features are available in the MyPersonality dataset. However, there are differences for the research carried out [26], use of a dictionary called Structured Programming for Linguistic Cue Extraction (SPLICE) as a method for performing feature extraction. By using the dictionary features such as positive or negative evaluation of the speaker, a value for the complexity and readability of a text can be generated. After that, the collected features will be compared with the two approaches using machine learning algorithms such as Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and deep learning architecture such as Convolutional Neural Network (CNN). However, the resulting performance is still low in several personality models, namely in the range of 60%–75% accuracy score. This is caused by due to small number of dataset used in this study to capture much more contextual information in creating generalized model.

Meanwhile, another study using Twitter dataset in Bahasa which were carried out by [3, 19, 31] used a different algorithm in building the personality prediction model. In feature extraction methods, the researcher was assessing the tendency user choice of words by using n-gram and LIWC. The implementation of close vocabulary method such as Term Frequency-Inverse Document Frequency (TF-IDF) is also applied in this study to show relationship between the main keywords discuss in their social media status data with their personality. This method used to filter out the least important words in the document and select the main topic in the sentences [9]. Another reliable open vocabulary feature extraction method called National Research Council (NRC) emotion lexicon

database was also introduced in the previous study. This corpus created by National Research Council Canada with about 14,000 words in English along with the association of these word associations with eight common emotions, namely anger, fear, anticipation, trust, surprise, sadness, joy, and disgust and the sentiments of each of these words which can be positive or negative [15]. Moreover, to apply such open vocabulary feature extraction method, the dataset is translated into English first before the feature extraction method is carried out. The algorithm used in these experiments is the ensemble method approach, namely stacking, boosting, and bagging, resulting in increased accuracy from each previous experiment. From this experiment a high accuracy value of around 97.9% is generated using this Twitter dataset. However, the researcher state that there are biases in the result due to the extremely small size of the dataset after sampling. The approach used can result in removing the contextual meaning of the sentence contained in the social media data.

In terms of the latest technology, the use of deep learning has been widely applied to improve performance in predicting a person's personality. As in the experiment conducted by [10, 17, 23, 41] using another dataset, namely personality Café, where there are differences of personality modeling, called Myers Briggs Type Indicator (MBTI) approach. The deep learning architectures such as Long Short-Term Memory (LSTM) and Transformer capable in producing a high performance of MBTI personality modeling with the maximum accuracy of 86.9%. Moreover, the use of pre-trained embedding has begun to be widely used in research to detect personality traits whereas in research [20, 22] use pre-trained model such as BERT and RoBERTa as feature embedding to create the model for Big Five personality model. The result from this research shows 67% and 60% accuracy. The obstacle of the two studies lies in the small amount of data so that additional data is needed for further research development. On the other hand, the use of pre-trained models is used to solve other NLP problems such as text based emoticon classification [2] and toxic comment classification [30] using RoBERTa and XLNet shows improvement in accuracy when added with other NLP Features such as TF-IDF and sentiment analysis.

Compared to all these approaches, this research has concentrated to capture personality of a person from multiple social media data Facebook and Twitter through a combination of deep learning architectures with model averaging. Researcher also used NLP features as additional features to deep learning architecture, obtained from psycho-linguistic and basic linguistic features.

Methodology

By focusing on utilizing social media data Facebook and Twitter this research will be carried out in three stages, namely initiation, model development, and model evaluation. The details related to each stage can be seen in Fig. 1.

At the initiation stage, data collection was carried out to increase the amount of Twitter data that had been collected from previous studies [3, 19, 27]. Twitter data that has been collected manually will be annotated with the help of a psychological expert to define the personality of each Twitter user. On the other hand, the Facebook dataset will use an open-source dataset called MyPersonality.

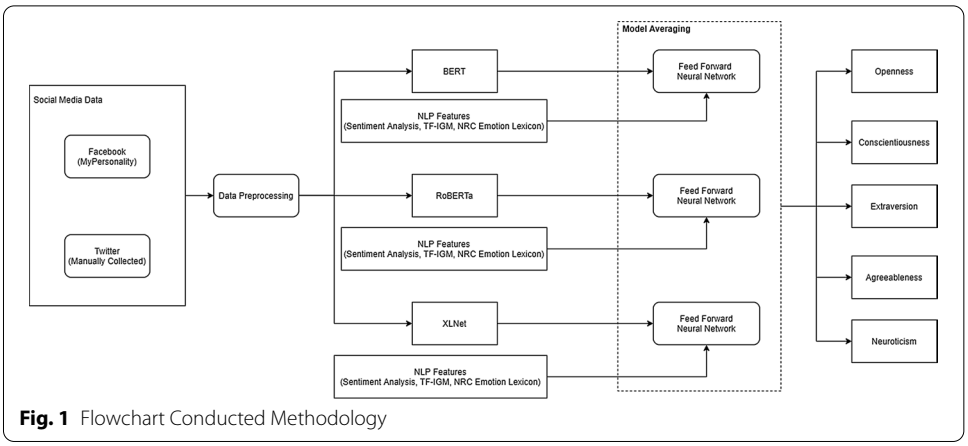

Fig. 1 Flowchart Conducted Methodology

Table 1 Big five personality traits characteristic [1]

	Openness	Conscientiousness	Extroversion	Agreeableness	Neuroticism
High	Creative, imaginative, abstract, curious	Discipline, obey, planner, ambitious	Communicative, friendly, assertive, active	Trusted, honest, humble, sympathetic	Anxious, nervous, worry, emotional
Low	Conservative, conventional, ordinary, usual	Lazy, easily give up, no purpose, unorganized	Alone, quiet, passive, unemotional	Critical, suspicious, stingy, grumpy	Quiet, emotionally controlled, comfortable, self-controlled

Each dataset used in this study uses the Big Five personality traits modeling approach to classify a person's personality. Each of these personalities is related to several characteristics that a person will have. Table 1 describes some of the characteristics of each dimension.

A person can have two types of values in each dimension. If someone who has a high personality dimension will be represented by the number one, and other will be represented by the number zero in the dataset. This label will later become a variable predictor in the model to be built. Furthermore, all data that have been collected will be preprocessed separately due to differences in the language used in each dataset. The results of the preprocessed data will be carried out by the feature extraction and feature selection process before entering the model building stage. Each personality based on five personality traits will be made a model that aims to predict each personality.

Data

The first dataset called the MyPersonality dataset, which consists of 250 users with a total of 9917 statuses. This dataset is collected through a Facebook app in 2007, allowing users to participate in psychological research by filling in the personality questionnaire [7]. The second dataset is an expanded dataset from previous research done by [27], which is a manually collected twitter data in Bahasa Indonesia. In this extended version, the dataset is added with new manually collected data resulting in a total of 502

Table 2 Facebook Dataset Distribution

Dataset	Facebook (MyPersonality)					
Type	Train		Test		Validation	
Label	No	Yes	No	Yes	No	Yes
Openness	1779	5133	381	1100	382	1100
Conscientiousness	3741	3171	801	680	802	680
Extraversion	3975	2937	852	629	852	630
Agreeableness	3235	3677	693	788	694	788
Neuroticism	4329	2583	928	553	928	554

Table 3 Twitter dataset distribution

Dataset	Twitter (manually collected)					
Type	Train		Test		Validation	
Label	No	Yes	No	Yes	No	Yes
Openness	14600	17511	3128	3753	3129	3753
Conscientiousness	23666	8445	5072	1809	5072	1810
Extraversion	9210	22901	1974	4907	1974	4908
Agreeableness	14348	17763	3074	3807	3075	3807
Neuroticism	17712	14399	3796	3085	3796	3086

users with 46,238 statuses. The addition of this twitter data was collected using the Twitter API. The same as previous research, the collected data will be annotated by psychology expert.

In addition, all datasets will be separated into three types, namely training set, test set, and validation set where the ratio of the distribution is 70% training set and 15% test and 15% validation. The data distribution for each dataset can be seen in Tables 2 and 3.

Preprocessing

All datasets will be preprocessed before feature extraction is carried out. The main purpose of preprocessing is to maximize the extracted features, hence more contextual features generated and normalized both datasets, since Twitter dataset written in Bahasa while Facebook dataset in English. The flow of the initial processes for both datasets is illustrated in Fig. 2.

In general, the two datasets are carried out the same preprocess. All data that has been collected will be removed from the use of URLs, symbols, and emoticons contained in social media status. Next, expanding a contraction in the sentence such as the use of I've become I have. After that, each sentence will be normalized by changing it to lowercase. Furthermore, any stopwords and clitics will be removed to prevent ambiguities. This list of words then will be processed using stemming function to normalize words by removing affixes to make sure that the resulting form is a known word in a dictionary. This preprocessing is carried out using the help of the NLTK library, which provides several linguistic functions to assist in cleansing social media status data such as tokenization, stemming, and stopwords dictionary. However, there will be an additional step during Twitter data preprocess, which is the translation process from Bahasa to English. In this

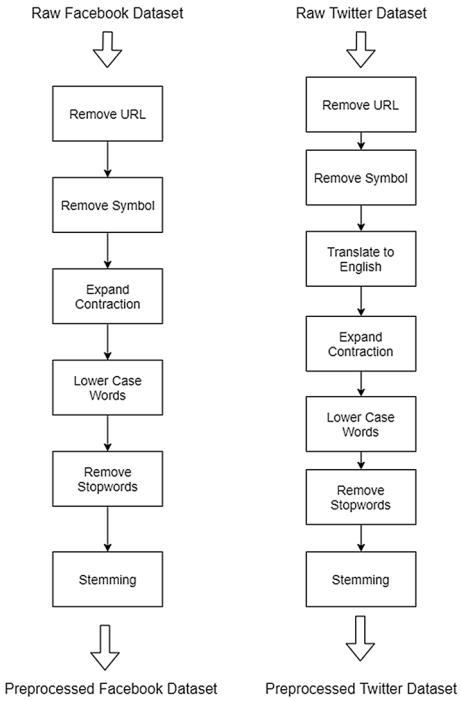

Fig. 2 Preprocessing stage

research this process is done by using Google translate API in translating Twitter status data.

Feature extraction

Researchers have recognized a novel combination of features which are profoundly compelling in aggression classification when applied in addition to the features obtained from the deep learning classifier at the classification layer. In this study, the researchers divide the feature extraction method into two types, pre-trained model features and statistical features.

As mention before the use of pre-trained model include BERT, RoBERTa, and XLNet. These pre-trained models are different from language representation modeling in general, where this architecture is designed to do the initial modeling of two-way representations in the unlabeled text by combining the context of each token in sentences from left to right and from right to left on each layer [33]. For these predefined models to be able to extract the context from a sentence, several preparations must be made to meet the necessary requirements. The following Fig. 3 visualize the step of feature extraction process using pre-trained models.

First, an example of social media status will be added by using a special token at the beginning and end of the sentence, namely [CLS] (stands for classification) and [SEP] (stands for separation). The purpose of these tokens is to serve as an input representation for classification tasks and to separate a pair of input texts respectively. Next, each word in a sentence will be tokenized which later become a sequence of words token. The tokenization is done using a method called WordPiece tokenization. This

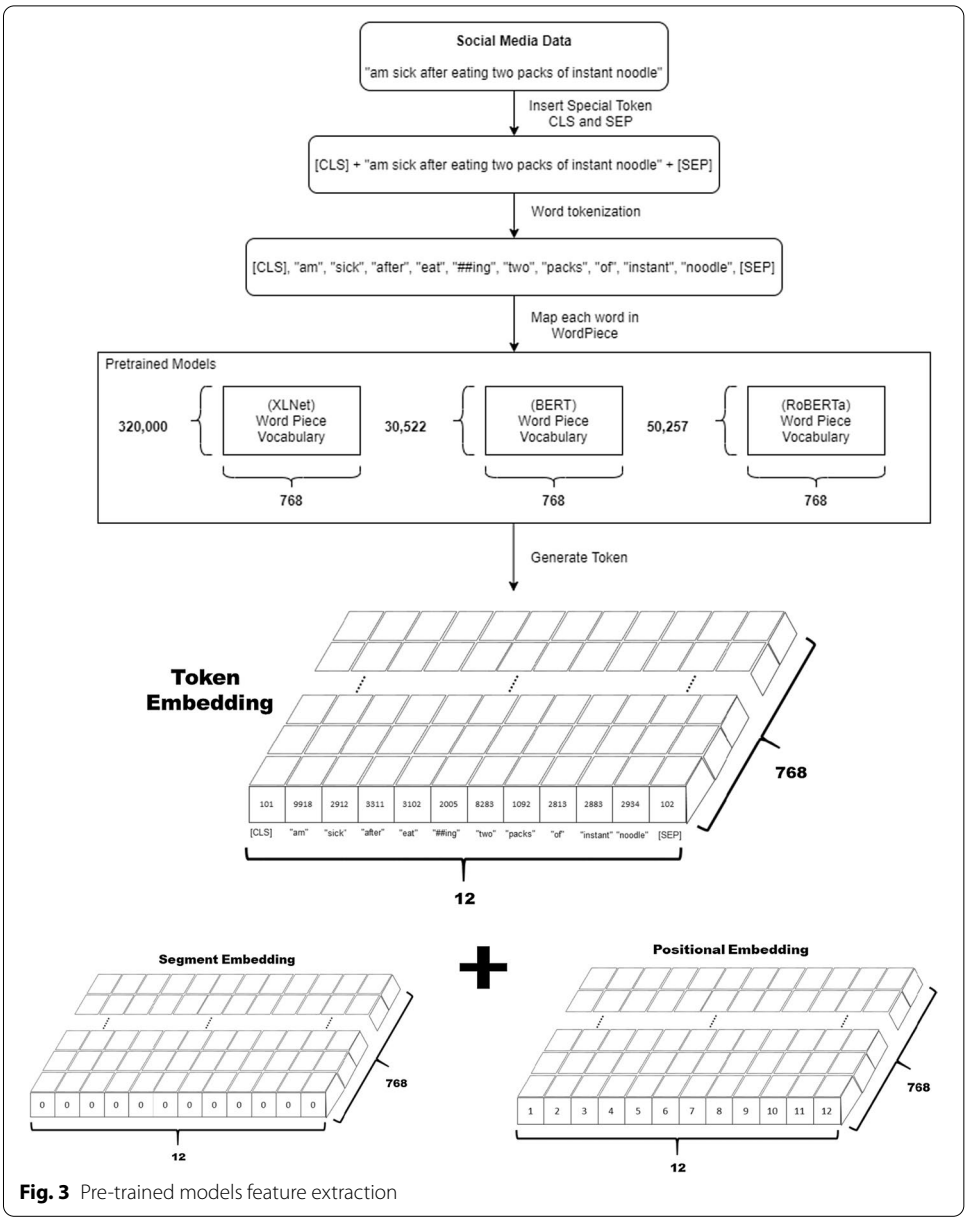

Fig. 3 Pre-trained models feature extraction

is a data-driven tokenization method that aims to achieve a balance between vocabulary size and out-of-vocab words. Each word that has been tokenized will be mapped with a WordPiece vocabulary. Each of pre-trained model has their own corpus dimension size for example BERT consist of 30,522 words, RoBERTa 50,257 words and XLNet 320,000 words. Each of these words represent in a 768 fixed dimension in a vector representation. In the example Fig. 3 the example social media status will be converted into a 12 tokens representation with each token consist of 768 lengths which is called token embedding. Before continuing to model building process this embedding will be added with another embedding layers called segment embedding and positional embedding to provide more contextual meaning to the model. The segment embeddings layer only has two vector representations. The first vector (index

0) is assigned to all tokens included in the first input while the last vector (index 1) is assigned to all tokens included in the second input. If the input consists of only one input sentence, then the segment embedded will only be the corresponding vector with index zero from the segment embeddings table. On the other hand, positional embedding layer is designed as a lookup table of sizes (n, 768) where n represent the number of length sentences. The first row is a vector representation of any word in the first position, the second row is a vector representation of each word in the second position and so on. The combination of those three embeddings called input embedding, act as a solution to overcome the limitations of architectures deep learning other such as the RNN which cannot capture sequence information, the combination of the three embeddings makes pre-trained model adaptable to NLP problems [12]. Table 4 describe list of pre-trained model along with maximum sequence length used to build the model, and references used to obtain them.

In terms of statistical features, this research uses different approaches compare to the previous research [37] which use TF-IDF as term weighting factor, instead TF-IGM is introduced in research. TF-IGM combine a new statistical model to precisely measure the weight of each class in text. The weight states the importance or word contribution to the class of documents. Furthermore, this method is able to separate label classes in a textual data, especially for data that has more than one label. Hence, this method is very suitable for use in personality prediction which allows a person to have more than one personality. The TF-IGM value can be calculated by looking for the TF value and the IGM value. TF represents the weight of a word, where how many words appear in a document. Meanwhile, IGM is useful for measuring the strength of a word in distinguishing between one class and another. The calculation of TF and IGM values can be described in the following formula:

Table 4 Pre-trained model features

Pre-trained Model	Description	Max sequence length used
BERT	The model has been trained using a very large data corpus that includes words from the Wikipedia site totaling 2.5 billion words and a dictionary containing 800 million words. The architecture consists of 12 encorder layers, 768 hidden units, and 12 attention heads. [13]	512
RoBERTa	RoBERTa is an extension of BERT, by adding a total of 16 GB of data from Wikipedia sources as well as additional data including the CommonCrawl News dataset (63 million articles, 76 GB), Web text corpus (38 GB) and Stories from Common Crawl (31 GB). The same architecture as BERT applied in this model. [27]	512
XLNet	XLNet is another development from BERT. This model introduces a permutation language modeling, where all tokens are predicted but in random order. This differs from the BERT language model where only 15% mask tokens are predicted. However, the number of layers, hidden units, and attention heads still the same as BERT. [40]	512
Total Pre-trained Model Features		1536

$$TF = \frac{Total\ appearance\ of\ a\ word\ in\ a\ document}{Total\ words\ in\ a\ document} \tag{1}$$

$$IGM = 1 + \lambda\left(\frac{Total\ appearance\ of\ a\ word\ in\ a\ document}{Total\ appearance\ of\ a\ word\ in\ each\ class}\right) \tag{2}$$

$$TF - IGM = TF * IGM \tag{3}$$

λ represent an adjustable coefficient, which use to keep the relative balance between the global and local factors in the weight of a term. Moreover, TF-IGM value will range from zero to 1. Each word in a document will be counted with a TF-IGM value then the words will be sorted according to the largest value. Words that have great value will be used as a feature for making classification models because it can be assumed that these words contain the important meaning of a document with a specific class label. Lastly, the use of semantic analysis and NRC emotion lexicon as correlating features in predicting characteristics of a person as were also used in this study. Both methods use the open vocabulary approach, which require a predefined corpus in finding the contextual feature from a text data. The Table 5 describe list of features, and references used to obtain them and the number of features.

Model prediction

Deep learning method has become more and more popular in recent periods, where several related studies use neural network architectures such as CNN and LSTM in making the best model for personality prediction systems. However, in this study, the multi model deep learning architecture was introduced by combining the statistical based text feature and a predefined model feature to improve the performance in predicting a personality of a person. In this research five classifiers will be made for this personality prediction system where each classifier represents the personality from the big five personality traits model. Figure 4 represent the model architecture that was build.

Each of input embedding extracted from the pre-trained model will be feed into a self-attention mechanism. Self-attention allows the models to associate each word in the input, to other words. It is also possible that the model learns that words structured in

Table 5 Statistical features

Feature name	Description	Feature count
TF-IGM	Statistical method to find how important a word is in a document influenced by the class label of a document. This method is used based on the research performance comparison between TF-IDF and TF-IGM in text classification [8]	100
Sentiment analysis	The percentage of positive, negative, and neutral in the social media status. The researcher used polarity sentiment analysis approach [35] to extract the weight for positive, negative & neutral class	3
NRC Lexicon Database	Contain 14000 set of words in English and the relation of each words with eight common emotions namely anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. [15]	8
Total statistical features		111

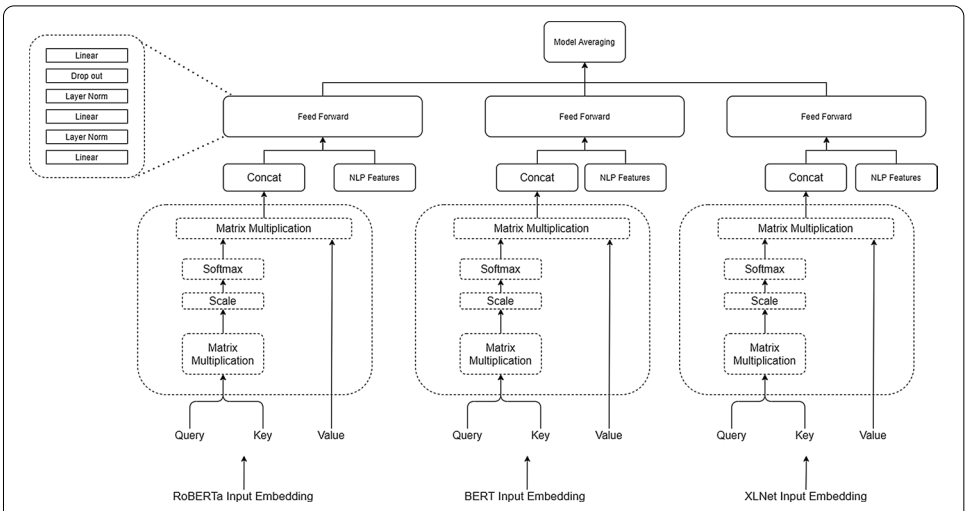

Fig. 4 Proposed Model Architecture

this pattern are typically a question so respond appropriately. To achieve self-attention, it feed the input into three distinct fully connected layers to create the query, key, and value vectors. The queries (Q) vector and keys (K) vector undergo a dot product matrix multiplication to produce a score matrix. The score matrix determines how much focus should a word be put on other words. So, each word will have a score that corresponds to other words in the time-step. The higher the score the more focus. This is how the queries are mapped to the keys. Next, the scores get scaled down by dividing with the square root of dimension query and keys (d_x). This process is to prevent exploding effect on the value hence, allowing for more stable gradient. After that, the softmax function is used to scaled down score to get attention weights and give an output in form of probability between zero and one. This function receives input of output matrix from scaled function (x_i) and sum of data inside the matrix (x_j). Applying this function makes the higher score get heighten and lower score depressed, therefore allowing the model to be more confident about which word to attend to. The softmax function and scaled down function denoted by the following formula.

$$Scaled(x) = \frac{\mathrm{QK}}{\sqrt{d_k}} \tag{4}$$

$$Softmax(x) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{5}$$

Finally, the attention weights will be multiplied with the value vector to get output vector. A higher softmax score will make the model learns that the higher value of the words means more important. Lower scores will eliminate irrelevant words. This final value will be concatenate to the original positional input embedding which is called a residual connection. The output of the residual connection will be combined with statistical NLP features with total of 623 features and inserted to a feed forward neural network. Inside each neural network will consist of three connected layers with alternating Rectified Linear Unit (ReLU) activation function and batch normalization in between them.

Moreover, a dropout function also applied in order to reduce overfitting and generalization error. Dropout deactivates the neurons randomly at each training step instead of training the data on the original network. In the next iteration of the training step, the hidden neurons which are deactivated by dropout changes because of its probabilistic behavior. Lastly, the output from the feedforward will be included in the averaging model function.

According to previous deep learning literature [21, 28], the unweighted averaging might be a reasonable ensemble for similar base learners of comparable performance. In this research the model averaging (unweighted) can be calculated by combining the softmax probabilities from three different classifications model. The mean class the probability is calculated as follow:

$$y_{i,k}^{*} = \frac{y_{i1,k} + y_{i2,k} + y_{i3,k}}{3} \forall k \in [1..K] \tag{6}$$

$$y = \arg\max\left(y_i^{*}, k\right) \tag{7}$$

where K is the number of classes, and y is the predicted label for a sentence. For loss function, a cross entropy loss denoted by the following formula was used.

$$Cross\ Entropy\ loss = -\left(ylog(p) + \left(1 - y\right)\log(1 - p)\right) \tag{8}$$

where y is the actual label value and p is the predicted personality of from a sentence. To maximize the performance of the model built, the parameter tuning process will be carried out. Grid search method will be used to perform repeated searches in finding optimal parameters that will produce the maximum level of predictive performance. Some of the parameters to be modified are the batch size, epoch, and learning rate.

Evaluation metric

The results of the model that have been created will be evaluated using several metric measurement approaches as follows:

a. F1 Measure

   Measurement metric of a model which combines the average values of precision and recall producing score by considering a classification error. These measurement metrics are best used when false negative and false positive values are important. In the prediction of personality, the false positive and false negative values are considered to reduce predictive errors because if the predictions are wrong, maybe someone can be placed incompatible with their personality.

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \tag{9}$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \tag{10}$$

$$F1Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{11}$$

b. Accuracy

Useful as a measure of the performance of a model, however this measurement focuses on the total data that is precisely predicted, namely true positive and true negative. This measurement is good for class distribution on balanced data. Based on previous research, many use this measurement as evaluation metric. Therefore, to compare the results of research with previous studies, this metric is used.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}. \tag{12}$$

Experiment

To compare the methodology that has been proposed previously, a comparison will be made by comparing to the other architecture with several feature extraction methods that are able to produce the best personality predictions in previous studies using the same big five personality model. Therefore, this research will be divided into several experimental scenarios where each scenario will use different algorithms along with feature extraction and feature selection method applied on two different datasets which are Facebook, and Twitter from to determine the performance of the proposed system. Table 4 shows the breakdown combination of each scenario (Table 6).

From the design scenario proposed, there are three types of model will be compare. First by using only pre-trained model (BERT, RoBERTa, XLNet). Second, with an addition of NLP statistical features which consist of NRC Lexicon Database, TF-IGM, and Sentiment Analysi. Lastly, the proposed architecture with model averaging from the three classifiers and addition of NLP statistical features. Each deep learning architectures will be tuned with different batch size and learning size on both social media datasets Facebook and Twitter.

Result and discussions

All predictive models that have been created will be evaluated using the accuracy and f1 measure metric approach. The results of the model evaluation can be seen in the table below. Table 7 is the result of the evaluation using the Facebook dataset, which shows that the highest accuracy produced from each trait dominated by the proposed model which used model averaging method and NLP statistical features. The first and second highest accuracy is produced in the Openness personality model with 86.17% accuracy and 0.912 f1 measure score, Neuroticism trait with an accuracy of 78.21% and f1 measure score 0.709. However, in terms of Agreeableness trait the highest accuracy and f1 measure score of other personality models is created by using the XLNet with addition of NLP features, which is found in Agreeableness personalities with accuracy values of 72.33% and 0.701. However, the resulting value differs slightly for about 0.74% and 0.011 from the proposed model in terms of both metrics. By looking at all the mean accuracy of the experiments on each algorithm, it was found that the proposed model architecture

Table 6 Experimental design scenarios

Scenario	System baseline	Batch size	Learning Rate
1	BERT	16	1.00E−05
2		16	3.00E−05
3		16	1.00E−05
4		16	3.00E−05
5		32	1.00E−05
6		32	3.00E−05
7		32	1.00E−05
8		32	3.00E−05
9	Roberta	16	1.00E−05
10		16	3.00E−05
11		16	1.00E−05
12		16	3.00E−05
13		32	1.00E−05
14		32	3.00E−05
15		32	1.00E−05
16		32	3.00E−05
17	XL Net	16	1.00E−05
18		16	3.00E−05
19		16	1.00E−05
20		16	3.00E−05
21		32	1.00E−05
22		32	3.00E−05
23		32	1.00E−05
24		32	3.00E−05
25	BERT + NLP Statistical Features	16	1.00E−05
26		16	3.00E−05
27		16	1.00E−05
28		16	3.00E−05
29		32	1.00E−05
30		32	3.00E−05
31		32	1.00E−05
32		32	3.00E−05
33	Roberta + NLP Statistical Features	16	1.00E−05
34		16	3.00E−05
35		16	1.00E−05
36		16	3.00E−05
37		32	1.00E−05
38		32	3.00E−05
39		32	1.00E−05
40		32	3.00E−05
41	XLNet + NLP Statistical Features	16	1.00E−05
42		16	3.00E−05
43		16	1.00E−05
44		16	3.00E−05
45		32	1.00E−05
46		32	3.00E−05
47		32	1.00E−05
48		32	3.00E−05

Table 6 (continued)

Scenario	System baseline	Batch size	Learning Rate
49	Proposed Method (Model averaging (BERT + ROBERTA + XLNet)) + NLP Statistical Features	16	1.00E−05
50		16	3.00E−05
51		16	1.00E−05
52		16	3.00E−05
53		32	1.00E−05
54		32	3.00E−05
55		32	1.00E−05
56		32	3.00E−05

Table 7 Personality prediction result using Facebook dataset

Traits	Metric	System baseline						
		BERT	**RoBERTa**	**XLnet**	**BERT + NLP features**	**RoBERTa + NLP features**	**XLNet + NLP features**	**Proposed model**
Openness	Accuracy	83.87%	81.11%	81.51%	84.68%	84.01%	84.35%	**86.17%**
	F1 Measure	0.897	0.878	0.879	0.902	0.898	0.899	**0.912**
Conscientiousness	Accuracy	68.96%	69.43%	69.64%	70.04%	69.70%	70.04%	**70.85%**
	F1 Measure	0.545	0.561	0.566	0.613	0.615	0.615	**0.652**
Extraversion	Accuracy	73.14%	72.33%	72.27%	74.76%	75.44%	75.51%	**76.92%**
	F1 Measure	0.742	0.705	0.707	0.736	0.739	0.735	**0.748**
Agreeableness	Accuracy	64.71%	65.32%	67.00%	70.11%	70.51%	**72.33%**	71.59%
	F1 Measure	0.617	0.620	0.641	0.680	0.676	**0.701**	0.690
Neuroticism	Accuracy	71.79%	71.05%	70.24%	73.08%	74.29%	73.75%	**78.21%**
	F1 Measure	0.637	0.621	0.621	0.658	0.675	0.667	**0.709**
Average	Accuracy	72.50%	71.85%	72.13%	74.53%	74.79%	75.20%	**77.34%**
	F1 Measure	0.688	0.677	0.683	0.718	0.720	0.723	**0.749**

The highest performance value resulting from each of the personalities listed in bold

has the highest average accuracy and f1 score, which is 77.34% and 0.749 for personality prediction system using Facebook dataset.

Furthermore, Table 8 defines the results of the evaluation using the Twitter dataset, which was collected manually. Similar to the previous results, the use of the proposed model architecture produces the best accuracy along with f1 measure score and dominates the highest performance from the five personality models. It shows the accuracy of 88.49% in the conscientiousness personality became the model with the highest accuracy, followed by extraversion personality, which was 81.17% and neuroticism was 75.08%. Differs from the previous results, the use of BERT with NLP features surpasses the results of the proposed model for the Agreeableness personality models with total accuracy of 72.33%. On the other hand, although the highest accuracy generated in conscientiousness trait is high, the f1 measure score gives a low result with value for only 0.652, this value is caused by the model tends to predict the low dominant trait rather

Table 8 Personality prediction result using Twitter dataset

Traits	Metric	System Baseline						
		BERT	RoBERTa	XLnet	BERT + NLP Features	RoBERTa + NLP Features	XLNet + NLP Features	Proposed Model
Openness	Accuracy	67.41%	66.38%	66.94%	69.28%	68.88%	69.85%	**70.85%**
	F1 Measure	0.702	0.691	0.697	0.723	0.718	0.729	**0.740**
Conscientiousness	Accuracy	81.76%	81.44%	81.24%	85.64%	85.86%	85.77%	**88.49%**
	F1 Measure	0.613	0.612	0.608	0.675	0.677	0.679	**0.736**
Extraversion	Accuracy	78.31%	77.80%	78.26%	78.99%	79.51%	79.50%	**81.17%**
	F1 Measure	0.863	0.860	0.862	0.867	0.871	0.871	**0.882**
Agreeableness	Accuracy	65.30%	65.14%	65.40%	**70.39%**	69.01%	68.44%	69.33%
	F1 Measure	0.694	0.695	0.694	**0.744**	0.731	0.723	0.734
Neuroticism	Accuracy	70.74%	70.47%	71.51%	73.57%	73.76%	73.99%	**75.08%**
	F1 Measure	0.631	0.632	0.643	0.673	0.677	0.679	**0.694**
Average	Accuracy	72.70%	72.25%	72.67%	75.57%	75.40%	75.51%	**77.34%**
	F1 Measure	0.701	0.698	0.701	0.736	0.735	0.736	**0.760**

The highest performance value resulting from each of the personalities listed in bold

Table 9 Final proposed model best parameters

Traits	Batch size	Learning rate
Openness	16	1.00E−05
Conscientiousness	32	1.00E−05
Extraversion	16	1.00E−05
Agreeableness	16	1.00E−05
Neuroticism	32	3.00E−05

than the high dominant trait therefore causing affect in precision and recall value. However, the proposed model architecture still give the highest results in an average accuracy and f1 score across other algorithms with a value of 77.34% and 0.760.

Next, Table 9 shows the results of the combination of parameters from the proposed deep learning model architecture which shows the best accuracy and f1 measure in performance. The model parameter tuning was carried out using tenfold cross validation. In determining the batch size and learning rate, the validation data mentioned in the previous section are used. From the result it is shown that for the model used to predict Openness, Extraversion, and Agreeableness required the batch size which is 16 while the other remaining traits required 32 batch size to get the optimum result. As for learning rate, all traits except Neuroticism use 3.00E−05 for the optimum performance. While the rest used 1.00E−5 for the optimum performance.

Finally, Tables 10 and 11 represent the comparative experimental results for the proposed method in this paper with respect to the state-of-the-art. To compare the overall system performance, some research uses the average accuracy and the other average f1-measure. The top 4 models given in Tables 10 and 11 are the best performing models for Facebook dataset and Twitter dataset with Big Five personality model as the label

Table 10 Comparison Model Performance (Facebook Dataset)

Facebook		
System	**Average accuracy**	**Average F1**
Tandera et al. [37]	70.40%	–
Zheng and Wu [42]	–	0.71
Tadesse et al. [36]	74.20%	–
Yuan et al. [41]	70.00%	–
Experiment model		
Scenario 1–8	72.50%	0.688
Scenario 9–16	71.85%	0.677
Scenario 17–24	72.13%	0.683
Scenario 25–32	74.53%	0.718
Scenario 33–40	74.79%	0.720
Scenario 41–48	75.20%	0.723
Scenario 49–56	**76.75%**	**0.742**

The highest values for the average accuracy and f1-measure for each personality model are shown in bold

Table 11 Comparison model performance (Twitter Dataset)

Twitter		
system	**Average accuracy**	**Average F1**
Pratama and Sarno [34]	65.00%	–
Ong et al. [32]	74.23%	–
Ong et al. [31]	70.50%	–
Ergu I [14]	75.7%	–
Experiment model		
Scenario 1–8	72.70%	0.701
Scenario 9–16	72.25%	0.698
Scenario 17–24	72.67%	0.701
Scenario 25–32	75.57%	0.736
Scenario 33–40	75.40%	0.735
Scenario 41–48	75.51%	0.736
Scenario 49–56	**76.98%**	**0.757**

The highest values for the average accuracy and f1-measure for each personality model are shown in bold

respectively. By analyzing these values, it can be concluded that the proposed deep learning architecture is able to provide the best model performance for in terms of the overall average performance of accuracy and f1-measure among all of the approaches. Moreover, the results also state that all the classifier with NLP features performs better compare to the individual pre-trained model features. This means providing NLP features will increase the model performance in predicting personality traits.

Conclusion

This research shows the comparison of different feature extraction method along with different algorithm approach in building personality prediction system for multiple social media data sources. Through this experiment, the proposed deep learning architecture approach with BERT, RoBERTa, XLNet as pre-trained language model, NLP

statistical features and model averaging outperform on most personality model builds by producing the highest accuracy of 86.17% and f1 measure score 0.912 on Facebook dataset and 88.49% accuracy and 0.882 f1 measure score on the Twitter dataset. Moreover, an addition of NLP statistical features such as TF-IGM, sentiment analysis, and NRC lexicon database contributed significantly to the personality prediction system on both datasets, since it can increase the model performance compare to only using pre-trained model as extraction features.

Future development of this experiment may utilize the use of larger training and testing dataset. Furthermore, another comparison approaches such as implementing another pre-trained model such as ALBERT which is A Lite BERT for Self-supervised Learning of Language Representation, DistilBERT, and BigBird may also be a possible candidate to increase accuracy in the personality prediction system.

Abbreviations

ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations; API: Application Programming Interface; BERT: Bidirectional Encoder from Transformer; CNN: Convolutional Neural Network; LDA: Linear Discriminant Analysis; LIWC: Linguistic Inquiry and Word Count; LSTM: Long Short Term Memory; MBTI: Myers Briggs Type Indicator; NLTK: Natural Language Toolkit; NLP: Natural Language Processing; NRC: National Research Council; RF: Random Forest; RNN: Recurrent Neural Network; ROBERTA: A Robustly Optimized BERT Pretraining Approach; SNA: Social Network Analysis; SPLICE: Structured Programming for Linguistic Cue Extraction; SVM: Support Vector Machine; TF-IDF: Term Frequency-Inverse Document Frequency; URL: Uniform Resource Locator; XGBoost: Extreme Gradient Boosting; XLNET: Generalized Autoregressive Pretraining for Language Understanding.

Acknowledgements

We would like to thank Bina Nusantara University for grant "Penelitian Internasional Binus" year 2020 numbered 080/VR.RTT/VIII/2020 which support our research. Supports from School of Computer Science, Bina Nusantara University and Universiti Malaysia Pahang for supporting all experiments in this research.

Authors' contributions

HC contributed as the research principal in this work as well as the technical issues. DS and AC advise all process for this work. Regarding the manuscript, HC, DS, AC and KZZ wrote and revised the manuscript. All authors read and approved the final manuscript.

Authors' information

Hans Christian is graduate student of Computer Science from Bina Nusantara University, Indonesia. He is working as data scientist at financial industry in Indonesia. His research interest includes artificial intelligence, machine learning, deep learning, natural language processing, and linguistics.

Derwin Suhartono is faculty member of Bina Nusantara University, Indonesia. He got his PhD degree in computer science from Universitas Indonesia in 2018. His research fields are natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a national scientific association in Indonesia. He has his professional memberships in ACM, INSTICC, and IACT. He also takes role as reviewer in several international conferences and journals.

Andry Chowanda is a Computer Science lecturer in Bina Nusantara University, Indonesia. He received his Ph.D. degree in Computer Science from The University of Nottingham UK, a master degree in Business Management from BINUS Business School ID, and a bachelor degree in Computer Science from BINUS University ID. His research is in agent architecture and Machine (and Deep) Learning. His work mainly on how to model an agent that has capability to sense and perceive the environment and react based on the perceived data in addition to the ability of building a social relationship with the user overtime. In addition, Andry is also interested in serious game and gamification design.

Kamal Z. Zamli received the degree in electrical engineering from the Worcester Polytechnic Institute, Worcester, MA, USA, in 1992, the M.Sc. degree in real-time software engineering from Universiti Teknologi Malaysia, in 2000, and the Ph.D. degree in software engineering from the University of Newcastle upon Tyne, U.K., in 2003. His research interests include search-based software engineering and computational intelligence.

Funding

All of this works is fully supported by Bina Nusantara University research grant namely PIB (***Penelitian Internasional Binus***) numbered 080/VR.RTT/VIII/2020.

Availability of data and materials

The datasets for this study are available on request to the corresponding author.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Computer Science Department, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia. [2]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia. [3]Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, 26600 Pahang, Malaysia.

Received: 22 January 2021 Accepted: 3 May 2021
Published online: 17 May 2021

References

1. Abood N. Big five traits: a critical review. Gadjah Mada Int J Business. 2019;21(2):159–86. https://doi.org/10.22146/gamaijb.34931.
2. Acheampong FA, Nunoo-Mensah H, Chen W. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif Intell Rev. 2021. https://doi.org/10.1007/s10462-021-09958-2.
3. Adi GYNN, Tandio MH, Ong V, Suhartono D. Optimization for automatic personality recognition on Twitter in Bahasa Indonesia. Procedia Comp Sci. 2018;135:473–80. https://doi.org/10.1016/j.procs.2018.08.199.
4. Alam F, Stepanov EA, Riccardi G. Personality traits recognition on social network—Facebook. AAAI Workshop—Technical Report, WS-13-01, 2013. pp 6–9.
5. Aung ZMM, Myint PH. Personality prediction based on content of facebook users: a literature review. Proceedings - 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2019; 2019. pp. 34–38. https://doi.org/10.1109/SNPD.2019.8935692.
6. Ben-Porat O, Hirsch S, Kuchy L, Elad G, Reichart R, Tennenholtz M. Predicting strategic behavior from free text. J Artif Intell Res. 2020;68:413–45. https://doi.org/10.1613/JAIR.1.11849.
7. Bin Tareaf R, Berger P, Hennig P, Meinel C. Cross-platform personality exploration system for online social networks: Facebook vs. Twitter Web Intell. 2020;18(1):35–51. https://doi.org/10.3233/WEB-200427.
8. Carvalho F, Guedesa GP. TF-IDFC-RF: a novel supervised term weighting scheme. ArXiv. 2020.
9. Christian H, Agus MP, Suhartono D. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). ComTech Comp Math Eng Appl. 2016;7(4):285. https://doi.org/10.21512/comtech.v7i4.3746.
10. Cui B (n.d.). Survey analysis of machine learning methods for natural language processing for MBTI Personality Type Prediction. http://cs229.stanford.edu/proj2017/final-reports/5242471.pdf.
11. Dalvi-Esfahani M, Niknafs A, Alaedini Z, Barati Ahmadabadi H, Kuss DJ, Ramayah T. Social Media Addiction and Empathy: Moderating impact of personality traits among high school students. Telematics Inform. 2020. https://doi.org/10.1016/j.tele.2020.101516.
12. Dandannavar PS, Mangalwede SR, Kulkarni PM. Social media text—a source for personality prediction. Proc Int Conference Comput Tech Electronics Mech Syst CTEMS. 2018;2018:62–5. https://doi.org/10.1109/CTEMS.2018.8769304.
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 2019. pp. 4171–4186.
14. Ergu İ. Twitter Verisi ve Makine Ö ğ renmesi Modelleriyle Ki ş ilik Tahminleme Predicting Personality with Twitter Data and Machine Learning Models. 1. 2019.
15. Farnadi G, Sushmita S, Sitaraman G, Ton N, De Cock M, Davalos S. A multivariate regression approach to personality impression recognition of vloggers. WCPR 2014 - Proceedings of the 2014 Workshop on Computational Personality Recognition, Workshop of MM 2014, 1–6. 2014. https://doi.org/10.1145/2659522.2659526.
16. Han S, Huang H, Tang Y. Knowledge of words: An interpretable approach for personality recognition from social media. Knowl-Based Syst. 2020;194:105550. https://doi.org/10.1016/j.knosys.2020.105550.
17. Hernandez and Knight. (n.d.). Predicting MBTI from text.
18. Howlader P, Pal KK, Cuzzocrea A, Kumar SDM. Predicting facebook-users' personality based on status and linguistic features via flexible regression analysis techniques. Proc ACM Symposium Appl Comput. 2018. https://doi.org/10.1145/3167132.3167166.
19. Jeremy NH, Prasetyo C, Suhartono D. Identifying personality traits for Indonesian user from twitter dataset. Int J Fuzzy Logic Intell Syst. 2019;19(4):283–9. https://doi.org/10.5391/IJFIS.2019.19.4.283.
20. Jiang H, Zhang X, Choi JD. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings. ArXiv, 2019. pp. 2–4.
21. Ju C, Laan MJ, Van Der (n.d.). The relative performance of ensemble methods with deep convolutional neural networks for image classification. pp. 1–20.

22. Kazameini A, Fatehi S, Mehta Y, Eetemadi S, Cambria E, Computational G, Unit N. Personality Trait Detection Using Bagged SVM over BERT Word Embedding Ensembles. 2020. pp. 1–4.
23. Keh SS, Cheng I-T. Myers-Briggs personality classification and personality-specific language generation using pre-trained language models. July. 2019. http://arxiv.org/abs/1907.06333.
24. Khurana D, Koli A, Khatter K, Singh S. Natural Language Processing : State of The Art , Current Trends and Challenges Natural Language Processing : State of The Art , Current Trends and Challenges Department of Computer Science and Engineering Manav Rachna International University , Faridabad-. ArXiv Preprint ArXiv, August 2017. 2018.
25. Kircaburun K, Alhabash S, Tosuntaş ŞB, Griffiths MD. Uses and gratifications of problematic social media use among university students: a simultaneous examination of the big five of personality traits, social media platforms, and social media use motives. Int J Ment Heal Addict. 2020;18(3):525–47. https://doi.org/10.1007/s11469-018-9940-6.
26 Lim HS, Bouchacourt L, Brown-Devlin N. Nonprofit organization advertising on social media: the role of personality, advertizing appeals, and bandwagon effects. J Consumer Behav. 2020. https://doi.org/10.1002/cb.1898.
27. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv; 2019. 1.
28. Lynn VE, Balasubramanian N, Schwartz HA. Hierarchical modeling for user personality prediction: the role of message-level attention. 2020. 5306–5316.
29. Marouf AA, Hasan MK, Mahmud H. Comparative analysis of feature selection algorithms for computational personality prediction from social media. IEEE Trans Comput Social Syst. 2020;7(3):587–99. https://doi.org/10.1109/TCSS.2020.2966910.
30 Maslej-kreš V, Sarnovský M, Butka P. Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. Appl Sci. 2020. https://doi.org/10.3390/app10238631.
31. Ong V, Rahmanto ADS, Williem W, Suhartono D, Nugroho AE, Andangsari EW, Suprayogi MN. Personality prediction based on Twitter information in Bahasa Indonesia. Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, 11; 2017. pp. 367–372. https://doi.org/10.15439/2017F359
32. Ong V, Rahmanto ADS, Williem, & Suhartono, D. . Exploring personality prediction from text on social media: a literature review. Internetworking Indonesia J. 2017;9(1):65–70.
33. Peters ME, Neumann M, Zettlemoyer L, Yih WT. Dissecting contextual word embeddings: Architecture and representation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018; 2020. pp. 1499–1509. https://doi.org/10.18653/v1/d18-1179.
34. Pratama BY, Sarno R. Personality classification based on Twitter text using Naive Bayes, KNN and SVM. Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015; 2016. pp. 170–174. https://doi.org/10.1109/ICODSE.2015.7436992.
35. Redhu S. Sentiment analysis using text mining: a review. Int J Data Sci Technol. 2018;4(2):49. https://doi.org/10.11648/j.ijdst.20180402.12.
36. Tadesse MM, Lin H, Xu B, Yang L. Personality predictions based on user behavior on the Facebook social media platform. IEEE Access. 2018;6(2016):61959–69. https://doi.org/10.1109/ACCESS.2018.2876502.
37. Tandera T, Hendro S, D., Wongso, R., & Prasetio, Y. L. . Personality prediction system from facebook users. Procedia Comp Sci. 2017;116:604–11. https://doi.org/10.1016/j.procs.2017.10.016.
38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem(Nips), 2017. pp. 5999–6009.
39. Violino B. Social media trends. Association for Computing Machinery. Commun ACM. 2020;54(2):17.
40. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. ArXiv, NeurIPS; 2019. pp. 1–18.
41. Yuan C, Wu J, Li H, Wang L. Personality recognition based on user generated content. 2018 15th International Conference on Service Systems and Service Management, ICSSSM 2018; 2018. pp. 1–6. https://doi.org/10.1109/ICSSSM.2018.8465006
42. Zheng H, Wu C. Predicting personality using facebook status based on semi-supervised learning. ACM Int Conference Proc Series, Part. 2019;F1481:59–64. https://doi.org/10.1145/3318299.3318363.

Publisher's Note