IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 13, 2018, accepted September 29, 2018, date of publication October 17, 2018, date of current version November 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2876502

Personality Predictions Based on User Behavior on the Facebook Social Media Platform

MICHAEL M. TADESSE, HONGFEI LIN, BO XU, AND LIANG YANG
Dalian University of Technology, Dalian 116024, China

Corresponding author: Hongfei Lin (hflin@dlut.edu.cn)

This work was supported in part by the Grant from the Natural Science Foundation of China under Grant 61632011, Grant 61572102, Grant 61402075, Grant 61602078, and Grant 61572098, in part by the Ministry of Education Humanities and Social Science Project under Grant 16YJCZH12, in part by the Fundamental Research Funds for the Central Universities under Grant DUT18ZD102, and in part by the National Key Research Development Program of China under Grant 2016YFB1001103.

ABSTRACT With the development of social networks, a large variety of approaches have been developed to define users' personalities based on their social activities and language use habits. Particular approaches differ with regard to different machine learning algorithms, data sources, and feature sets. The goal of this paper is to investigate the predictability of the personality traits of Facebook users based on different features and measures of the Big 5 model. We examine the presence of structures of social networks and linguistic features relative to personality interactions using the myPersonality project data set. We analyze and compare four machine learning models and perform the correlation between each of the feature sets and personality traits. The results for the prediction accuracy show that even if tested under the same data set, the personality prediction system built on the XGBoost classifier outperforms the average baseline for all the feature sets, with a highest prediction accuracy of 74.2%. The best prediction performance was reached for the extraversion trait by using the individual social network analysis features set, which achieved a higher personality prediction accuracy of 78.6%.

INDEX TERMS Big 5, feature analysis, predicting personality, social behavior, social networks.

I. INTRODUCTION

In recent years, social media such as Facebook, Twitter and Weibo have become some of the most popular destinations for internet users. These users' activities on social networks provide a great platform for researchers to study and understand their online behaviors, preferences and personalities. Different personalities are related to the formation of different social relations and interaction behaviours on status profiles or preferences. Our study predicts personality based on users' social behavior and their language-use habits on Facebook social media platform. First, we choose most beneficial features for each personality dimensions and successfully predict the user's personality. Next, we propose a method to design and implement one category of Social Network Analysis (SNA) features and two categories of linguistic features such as Linguistic Inquiry and Word Count (LIWC) and Structured Programming for Linguistic Cue Extraction (SPLICE) based on the myPersonality dataset. We explore the correlations between each of the feature sets and personality traits. In the case of social features, we focus on several classes of structural network properties, namely, networksize, betweenness, density, brokerage and transitivity measures as well as their relationships to particular traits. We then investigate the predictive power of the features by predicting each personality trait. We examine the features with the highest correlations. Finally, we set machine learning algorithms, implement them in the prediction model to explore the degree to which we can predict personality traits from Facebook and compare the highest algorithm accuracies for the Big Five personality traits. For classification comparisons, we apply three different machine learning algorithms as baseline methods for the prediction model to the correlated features data. In our study, we exploit data collected by means of the Facebook social media platform. Targeting the automatic recognition of the Big 5, our research paper extends and merges the lines of research followed by Staiano *et al.* [1].

Our study has four specific contributions: first, to clarify the relationship between users personalities and their interactions behavior in social networks; second, to illustrate a higher potential of individual social network features for personality prediction by using XGBoost machine learning approach; third, to show the relationship between LIWC dictionaries and SNA features set; forth, to introduce the cases when a higher prediction performance can be achieved with SNA rather than linguistic features.

The rest of the paper is organized as follows. In section II, we discuss related work in personality prediction. In section III, we introduce and describe the properties of the Facebook dataset. In section IV, we conduct the methodology and data preprocessing followed by feature extraction and feature selection. The dataset is categorized into two groups, text features extraction and social interaction behavior analysis, which are followed by a feature selection that defines the standard feature selection method to predict the personality traits. The correlation results are analyzed in section V, and the results of the prediction accuracy with different prediction models using SNA, LIWC and SPLICE feature sets are shown in section VI. We conclude our study and suggest future work in section VII.

TABLE 1. Summary of previous personality prediction studies.

Dataset	Author	Features	Method	Best Results
Facebook	Tandera et al.	LIWC, SPLICE SNA	SVM	Accuracy 70.4%
	Schwartz et al.	n-grams, extracted topics	R (square root of coefficient determination)	R 0.42
	Farnadi et al.	LIWC, SNA , time-related feature, others	SVM	Precision 0.71
Tweets	Ong et al.	LIWC	Xgoost	Accuracy 97.9%
	Golbeck et al.	LIWC	ZeroR	MAE 0.118
	I. F. Iatan	LIWC	ANN	NRMSE 0.079
Blogger	T. Yarkoni	LIWC, n-grams	Spearman's rank correlation coefficient	P 0.32

II. BACKGROUND AND RELATED WORK

There is a growing number of research papers related to a user's behavior in social networks that has recently attracted more attention in the international research community. Personality recognition is studied by two main disciplines: computational linguistics and Social Network Analysis. From the area of computational linguistics, Pennebaker *et al.* [2] wrote a pioneering work dedicated to personality extraction from text. They examined words in a variety of domains such as diaries, college writing assignments and social psychology manuscripts to study personality related features with linguistic cues. Their results show that agreeable people tend to use more articles while introverts and those low in conscientiousness use more words signaling distinctions. Neurotics use more negative emotion words. Argamon *et al.* [3] classified neuroticism and extraversion using linguistic features such as function words, judgemental and appraisal expressions and modal verbs. Their results revealed that neuroticism is related to the use of functional lexical features, for instance appraisal lexical taxonomy, whereas the results for extraversion were less clear. Other studies linked neuroticism to irrational beliefs or poor coping efforts on well-being personality [4] oberlander and Nowson 2006 classified the extraversion, stability, agreeableness and conscientiousness of bloggers using the Naive Bayes prediction model as a learning algorithm using different sets of n-grams as features. Karney and Bradbury [5] examined correlations between the Big 5 personality traits, using LIWC and RMC as feature sets. While LIWC features included word classification such as positive emotions or anger, RMC features included results about word age of acquisition or word imageability. Using the corpus of Essays written by Pennebaker and King in 1999, Mairesse *et al.* developed a supervised system for personality recognition.

In Social Network Analysis, personality recognition extracted from network configuration and other extra-linguistic cues has an even shorter history. The impact of a user's social interaction behavior on personality was studied by Gosling *et al.* [9]. They examined personality traits from self-reported Facebook usage and observable profile information. All the users' features were based on statistical characteristics instead of psychological properties. Davis *et al.* [10] showed that people can judge others' personalities from their Facebook profiles. Golbeck *et al.* [11] predicted the personality of 279 users from Facebook using linguistic features such as word count and social network features such as friends count. Ross *et al.* [12] revealed that shyness is positively correlated with the time spent online and negatively correlated with the number of friends. Sumner *et al.* [13] found the correlation between users' personalities and their Facebook usage, posts content and emotion. Their result indicated that openness is positively associated with words expressing negative emotions, anger, taboo subjects, money, religion or death. Kalish and Robins [14] experimentally examined the effect of individuals' personality differences on their immediate network environment, focusing on ego networks which consist of a focal node or ego and the nodes to which the ego is directly connected (the so-called alters) and the ties, if any, among the alters. Their findings showed that psychological predispositions can explain the variance portions of egocentric network characteristics. Personality prediction based on language features has received much interest in prediction research [15]–[17] however, from a network perspective, the role of links in supporting personality relations is not yet well understood. The aim of our study is to examine the presence of a structure of social networks and linguistic features relative to personality interactions using the myPersonality project dataset [18].

The myPersonality dataset used in our study is a sample of personality scores on Facebook profile data. The data were collected by Schwartz *et al.* [16] by means of a Facebook application that implemented the Big 5 personality traits' test among other psychological tests. The application includes obtaining consent from the users to record their

TABLE 2. Overview of the big five personality traits [6]–[8].

Personality Trait	Characteristics
Openness (O)	From cautious/consistent to curious/inventive intellectual, polished, creative, independent, open-minded, imaginative, creative, curious, tolerant
Conscientiousness (C)	From careless/easy-going to organized/efficient reliable, consistent, self-disciplined, organized, hard working, has long-term goals, planner
Extraversion (E)	From solitary/reserved to outgoing/energetic, express positive emotions, excited, satisfied, friendly, seeks stimulation in the company of others, talkative
Agreeableness (A)	From cold/unkind to friendly/compassionate kind, concerned, truthful, good natured, trustful, cooperative, helpful, nurturing, optimistic
Neuroticism (N)	From secure/calm to unconfident/nervous angry, anxious, neurotic, upset, depressed, sensitive, moody

data and use it for various research purposes. For instance, Bachrach *et al.* [19] used the myPersonality dataset to find the relationship between users' activity behavior and personality. The results showed that agreeableness is positively correlated with the number of tags, whereas neuroticism has a significantly negative correlation with the number of friends. Farnadi *et al.* 2013 [20] studied the relationship between the emotions expressed through Facebook status updates and user's age, gender and personality. He found that users with openness have a tendency to be more emotional in their status posts than users with neuroticism. Cantador *et al.* 2013 [21] used the dataset to study the relationship between personality types and user preferences in multiple entertainment domains such as music, movies, TV shows and books. Recent studies conducted by Tandera *et al.* [22] used two Facebook datasets, one from myPersonality and the other one manually collected. They used word embedding and the features from LIWC and SPLICE to predict personality based on the Big 5 model to classify the traits. Using the support vector machine learning method on the myPersonality dataset, they achieved the highest prediction accuracy of 70.40%. Ong *et al.* [23] also predicted personality using XGBoost trained on 329 users of Twitter social media, summary of personality prediction mentioned in this study in Table 1.

In psychology, the theory based on the Big 5 factors is the most widely accepted model to describe the basic structure of human personality. The theory based on these factors is called five factor model (or the Big 5 model) and it is the most widely accepted model of personality. It provides a nomenclature and a conceptual framework that unifies much of the research findings in the psychology of individual differences and personality. It reduces the large number of personal adjectives into five main personality traits that form the acronym OCEAN [24], [25]. It was first studied in the 1990s when five factors or personality traits were established and has been used until the present time. According to Table 2, individuals in the Big 5 model vary in terms of the OCEAN, that is openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. It represents a complete set of traits that could capture personality differences [8].

III. DATASETS

To examine personality traits from social networks, we employed the myPersonality dataset as a case study.

TABLE 3. Distribution of personality traits.

Value	O	C	E	A	N
Yes	176	130	96	134	99
NO	74	120	154	116	151

Note: O, C, E, A, N refer to the distribution of personality scores on the Big Five traits in Facebook: openness, conscientiousness, extraversion, agreeableness and neuroticism.

We constructed the study with 250 users and 9917 status updates from the myPersonality sample. The dataset of Facebook users was labeled according to the Big 5 model. According to the distribution of personality types in Table 3, each user in the dataset had multiple posts gathered in one file [18].

with some of the dataset, we selected the users information, such as the user's social network structure, user's status, and text posts. The final dataset contains the Facebook statuses in raw text, author information, personality labels (scores and classes) and five social network measures of the users in inference with personality traits, that is networksize, betweenness, density, brokerage and transitivity.

IV. METHODOLOGY

Since it has become increasingly popular to use language in social media for predicting personality [26], there is a growing number of methodologies that use both linguistic and social network features of profiles and status updates to infer personality traits. The personality prediction framework in Fig. 1 consists of data pre-processing, feature extraction and feature selection followed by the machine learning process and prediction results.

A. DATA PRE-PROCESSING

The dataset obtained from myPersonality was pre-processed before it proceeded to the feature selection and training stage. To pre-process the dataset, we employed OpenNLP [27]. First, we used tokenization in order to separate the last word of each sentence with punctuation and an aggregation of the same words. Next, we removed URLs, symbols, names, spaces and lower cases. Since many of the words in LIWC

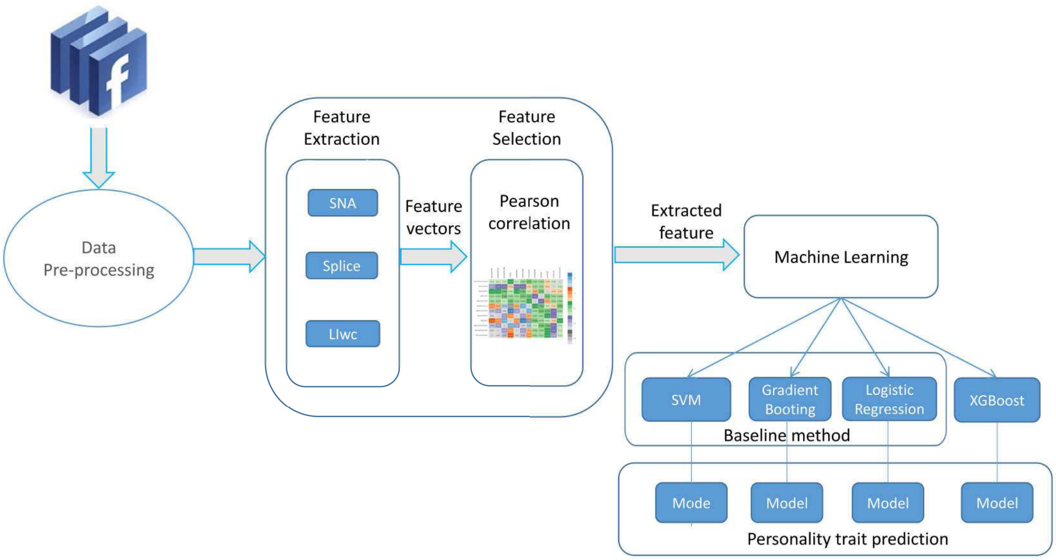


FIGURE 1. **Personality Prediction Framework.**

and SPLICE linguistic features share common stems, the relationship between personality and stemmed words could be negatively affected. For instance, in the case of tenses, such as present or past tense, verbs stemming would make it impossible to distinguish between particular tenses [28]. Hence, the correlation analysis in the pre-processing part of our experiment does not apply stemming, and all the words are left unstemmed.

B. FEATURES EXTRACTION

A user's behavior on social networks is mutually affected by the presence and behavior of other users. These interactions can have an impact on the transition of new information or behaviours through the groups. There are many potential applications for understanding how such behaviours arise and spread [29]. In our study, all the information from the dataset can be categorized into two groups. The first group is the text features extraction which reflects a user's language habits on Facebook and contains an expressions count and a topics count. To analyze the content of Facebook status texts, we use two dictionaries, namely, LIWC and SPLICE. The second group is the social interaction behavior analysis, which contains networksize, density, brokerage and transitivity. This information reflects a user's basic social network behavior on Facebook.

LIWC, or the Linguistic Inquiry and Word Count dictionary, is widely used in psychology studies [26]. In our study, we use it to extract 85 linguistic features from the texts including five subcategories such as standard counts (e.g., word count, words longer than six letters, number of prepositions), psychological processes (e.g., emotional, cognitive, sensory, social and emotional processes), relativity (e.g., words about time or tense verbs), personal concerns (e.g., occupation words such as job, majors, financial issues or health), and other linguistic dimensions (e.g., counts of various types of punctuation, swear words) [30]. For the text analysis, we chose LIWC2015 which is designed to analyze individual or multiple language files quickly and efficiently. In comparison to LIWC 2007 and LIWC 2001, it attempts to be transparent and flexible in its operation, allowing the user to explore word use in multiple ways [2].

SPLICE, or the Structured Programming for Linguistic Cue Extraction, is a newer dictionary developed in recent years. It is still going through the updating process and will be widely used for personality prediction tasks studies [31]. In our study, we use it to extract 74 linguistic features including cues that relate to the positive or negative self-evaluation of the speaker, complexity and readability scores.

SNA, or Social Network Analysis, is a technique that analyzes the social structure that emerges from the combination of relationships among members of a given population or a network of relationships and interactions with nodes (representing "actors" or people on whom relations act within the network) and ties (representing the relationships among these actors) [32], [33]. It is an approach for examining and quantifying the patterns of relationships that arise among interacting social entities, especially individuals. An explicit assumption of this approach is that indirect relationships (e.g., friends of friends) in social groups matter. According to a study by James and Christakis [34], happiness tends to be correlated in social networks. When a person is happy, close friends have a 25% higher chance of being happy, too. Furthermore, people at the centre of a social network tend to become happier in the future than those at the periphery. Clusters of happy and unhappy people were discerned within the studied networks with a reach of three degrees of separation. Person's happiness was associated with the level of happiness of their friends' friends' friends.

For our analysis we used features related to the social network of a user in inference with personality traits, that is, networksize, betweenness, density, brokerage and transitivity. Networksize refers to the number of nodes in a network, which reflects the quantity of the connections [35]. Betweenness indicates the number of shortest connected paths between pairs of individuals who are not connected to each other directly. For instance, an individual high in betweenness is critical for the flow of information among other individuals who do not know each other directly [36]–[38]. Density addresses the quality of interpersonal relations calculated as a proportion of a number of edges existing in the network relative to the number of maximum possible edges in the same network. High density networks cause a high diffusion between nodes in the information flow [39]. Brokerage is the number of incoming ties the individual receives from others. An individual is connected to people or clusters of people by being active within the network, maintaining many ties to be an efficient and important go-between to other vertices in the network [40]. Transitivity is based on "friends of my friends are also my friends" idea in which two or three individuals are directly connected with each other through a mutual neighbor. One of them is only accessible via another individual's ties and represents the frequency of interactions among the network's nodes or shows the social relation among the nodes [41], [42].

C. FEATURE SELECTION

Generally, there are two main reasons why feature selection is important for building a model. First, it reduces the high dimensionality of the dataset by removing the features not essential for training, improving the generalization of the model and reducing the training time. Second, the model gains a better understanding of the features and their relationships to the response features. Additionally, it improves the accuracy of the learning algorithms and reduces the processing requirements [20], [43], [44].

To measure the strength of the linear relationship between two variables and to examine features important for personality traits prediction, we used the Pearson correlation analysis, Eq. (1), as the standard feature selection method. Pearson correlation is a measure of the linear correlations between two variables, and we used it to predict the relationship between the personality scores and extracted features. For a pair of variables (x, y), the linear correlation coefficient r is given by the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are sample means given by the relations

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

n in the above equation represents the sample size, and x_i, and y_i describe the single samples indexed with i, where the value of r lies between -1 and 1 inclusive. If x and y are completely correlated, r takes either the value of 1 as a positive correlation or -1 as a negative correlation. If x and y are completely independent, r is zero [45], [46].

TABLE 4. Pearson correlation values between LWIC and personality traits.

	3rd person singular	Prepositions	Social processes words	Affective process words	Death
O	**0.16**	0.06	**0.12**	**0.14**	0.07
	1st person singular	Common verbs	Dictionary words	Netspeak	Perceptual processes related words
C	**-0.15**	**-0.13**	**0.11**	**-0.12**	**-0.14**
	2nd Person singular	Past tense verbs	Positive emotion	Agreement words	Achievements words
E	**0.15**	**0.14**	**0.1**	**0.13**	**0.19**
	Interrogative sentence	Biological processes	Sexual words	Social process words	Assent words
A	**0.17**	**0.11**	**0.19**	**-0.14**	0.09
	Positive emotions words	Anger	Affective process words	Social process word	Prepositions
N	**-0.12**	**0.15**	**0.12**	**-0.11**	**-0.14**

Note: O, C, E, A, N refer to the distribution of personality scores on the Big Five traits in Facebook: openness, conscientiousness, extraversion, agreeableness and neuroticism.

V. EXPERIMENTAL RESULTS

In this section of our paper, we first analyze the Pearson correlation coefficient between three feature sets and personality scores to quantify the importance of each feature. The results are shown in Table 4, Table 5 and Table 6. Due to the limitation in the text space, we only discuss the features that show a significant correlation ($p < 0.05$) between the features set value and personality score based on the r value. They are all bolded.

Several interesting observations were made through the correlation between the LIWC features and personality traits. Extraverted users tend to use 2nd (0.15) and 3rd person singular pronouns (0.16) and past tense verbs (0.14). They often update their statuses with dictionary words (0.16), social interaction words (0.14) and common adjectives (0.17). Concerning the length of the texts, extraverts prefer to write short messages as they negatively correlate with word

TABLE 5. Pearson correlation values between splice features and personality traits.

	Present tense	Imagery	Activation	SWN Negativity	Pleasantness
O	0.05	**0.13**	**0.12**	**0.13**	**0.15**
	Agreement ratio	Total submissiveness	Submissiveness ratio	Verbal words	SWN positivity
C	**-0.1**	**-0.12**	**-0.14**	-0.07	0.01
	I cando it	Average sentence length	Activation	Complexity composite	Pleasantness
E	**0.12**	-0.09	**0.1**	**-0.16**	**0.12**
	Complexity composite	Pausality	Num agreement	Num interjections	Question-count
A	**-0.12**	**-0.12**	0.09	0.06	0.04
	Complexity composite	Expressivity	Activation	PoSelfImage	Pleasantness
N	0.12	**0.11**	**-0.12**	-0.05	**-0.14**

Note: O, C, E, A, N refer to the distribution of personality scores on the Big Five traits in Facebook: openness, conscientiousness, extraversion, agreeableness and neuroticism.

TABLE 6. Pearson correlation values between SNA and personality traits.

Traits	Networksize	Betweenness	Density	Brokerage	Transitivity
O	0.02	0.04	0.05	0.04	-0.06
C	**0.14**	**0.11**	**-0.14**	**0.11**	-0.03
E	**0.31**	**0.25**	**-0.24**	**0.25**	-0.3
A	0.07	0.05	-0.08	0.05	**-0.15**
N	**-0.18**	-0.03	**0.11**	**-0.13**	**0.14**

Note: O, C, E, A, N refer to the distribution of personality scores on the Big Five traits in Facebook: openness, conscientiousness, extraversion, agreeableness and neuroticism.

length (−0.08) and sentence length (−0.09), suggesting that people who score high on extraversion tend to use fewer words and shorter sentences in their status posts. On the other hand, they also use more words indicating positive emotions (0.6), such as "love", "nice" or "sweet"; agreement (0.13), such as "OK", "yes" or "agree"; social interaction (0.12), such as "friend" or "family"; and interpersonal interaction, suggesting that the more extraverted individuals are, the more likely they are to talk about personal acquaintances [13] and achievements (0.19). Additionally, the results show that extraverts tend to update their statuses with more emotional words than neurotic users. This finding is supported by the work of [47].

Furthermore, neurotic users tend to update their statuses with word categories expressing negative emotion, such as anger (0.15), anxiety (0.08) or affective processes (0.13). They are less likely to use words indicating social interaction (−0.13) or positive emotions (−0.14). When talking to others, they often use "we" and "ours" (0.05), whereas extraverts prefer using "he" or "she" pronouns. Neurotics positively correlate with the lengths of words and sentences (0.06), indicating that people high in neuroticism will write more sentences linked to negative emotions, anxiety and irritability. This finding is also supported by [48]. Since neurotic and extraverted users positively correlate with the words connected with biological processes, they are more likely to share information about body and health on social networks.

Users high in openness have a tendency to favour high-frequency function words rather than content words. Their status posts contain more articles, prepositions (0.06) and personal pronouns such as "they" (0.13), "he", "she" or "him" (0.15). They use longer expressions in their sentences, have positive correlation for average word count (0.046), and words per sentence as well as words greater than six letters (0.087). Although they frequently update their statuses with dictionary words (0.15), social processes words (0.13), affective processes (0.14) and cues associated with perceptual processes, such as "listen" (0.18), and tentative words of certainty, such as "unsure", "uncertain" (−0.06) or "never" (−0.02) appear with a lower frequency. Our results support the work of Sumner *et al.* [13], suggesting that people with higher levels of openness may be more open to talking about potentially sensitive subjects. They often positively correlate with words expressing negative emotions (0.16), religion (0.18) or death (0.071).

Conscientious users negatively correlate with the words expressing negative emotions (−0.14), such as "hurt", "ugly" or "nasty" (−0.16). In contrast, they tend to talk less about unhappy subjects. They update their statuses with words describing perceptual processes (0.12) such as "see", "hear" or "feel" as well as the words surrounding social processes (0.12), indicating that highly conscientious people like to discuss with other people and often talk about the things they see or hear. Furthermore, conscientious users positively correlate with dictionary words (0.11), suggesting that they are more likely to use properly spelled words than informal words such as "btw", "lol", "thx" (−0.12) or typical cyber words [49]. Additionally, they are less likely to use verbs (−0.13) and 1st person singular pronouns (−0.15). Agreeable users are more likely to use interrogative sentences and question marks (0.17). They prefer using the "I" pronoun and words indicating biological processes such as body (0.11) or sexual (0.19). They are less likely to share their information about social processes (−0.14).

According to the correlation results between SPLICE and the Big 5 personality traits, extraverts, in comparison with neurotic users positively correlate with PosSelfImage words (0.08) and portray themselves positively in the text. Users high in openness highly correlate with imagery (0.13), which indicates that they are usually viewed as imaginative people with individual curiosity, open mindedness and willingness to explore new ideas [6].

People scoring high on openness and agreeableness update their statuses with activation words (0.12, 0.1). They use words expressing pleasantness (0.15, 0.12) and verbs in the present tense (0.05, 0.013). Users high in openness positively correlate with complexity composite words (0.12), while users high in agreeableness and extraverts are less likely to use them (−0.12 and −0.16, respectively). Conscientious users negatively correlate with AgreementRatio (−0.1), TotalSubmissiveness (−0.12) and SubmissivenessRatio (−0.14), indicating that these kinds of individuals are not easily manipulated by others and like to maintain their original plan. According to our results, there are some correlation similarities between the results of LIWC and SPLICE, as shown in Table 4 and Table 5 respectively. For instance, taking motivational words into consideration, extraverts highly correlate with ICanDoIt phrases (SPLICE 0.12) and achievement words such as ''win'', ''success'', or ''better'' (LIWC 0.19). Concerning word length in status posts, extraverts are less likely to use averageSentenceLength (SPLICE −0.09), averageWordLength (SPLICE −0.06) as well as word count LIWC −0.08), suggesting that extraverts tend to share information containing short sentences and fewer words. Conscientious users negatively correlate with the use of verbal words (SPLICE −0.037, LIWC −0.13) [6], which is supported by both dictionaries. Our results reveal that using different dictionaries can improve the prediction result.

Based on the correlation results for the social network features, we found that extraversion represents the highest correlated trait. Extraversion is a personality factor highly related to Facebook usage. These findings support the work of [9], [12], and [50] who claimed that extraverts have a positive relationship with social networks. In online and social media networks, they have the tendency to go online to seek out a new and exciting experiences [51]. Our results also support this finding. The results in Table 6 show that extraverts positively correlate with networksize (0.31), which reflects their tendency to have many friends. Opposed to neurotic users, however, their friends often do not know each other as they belong to different groups of people. This finding is supported by the negative results for transitivity (−0.3) as a connectivity feature in social networks. At the same time, the networks of their friends, tend to be sparser, as density shows a negative correlation (−0.24). Our results support the fact that extraverted individuals are sensitive to reward signals. They seek stimulation and participate in a wide variety of social activities. According to their socializing tendencies, they tend to have larger friendship networks. The number of missing connections among their contacts, however, is rather high.

Other negative correlation results for transitivity features were achieved for openness (−0.06), consciousness (−0.03) and agreeableness (−0.15). In other words, sociable individuals tend to be at the centre of large and loosely connected networks, which explains the negative effect of all four traits on transitivity. Neuroticism, however, had a positive correlation with transitivity (0.14), associated with its higher degree of nodes that are common neighbours that share the same experience.

Neuroticism interestingly showed the opposite correlation results for all the SNA features. In comparison to extraversion, neurotic users are negatively correlated with networksize (−0.18); they are identified as unpopular interaction partners in online discussion networks [52] and create small groups of friends. Due to the positive correlation with transitivity (0.14) and density (0.11), their friends, however, tend to know each other. Neurotic users report lower internet usages and information based activities which might be due to their higher level of anxiety, feelings of worry and insecurity among others [53]. While people high in extraversion are emotionally stable and tend to maintain persistent communications with their friends, neurotic users withdraw from communication with others, especially during times of stress and report less satisfaction with the support received from their friends on social networks [54], [55]. They rarely seek new experiences but are more likely to have self-efficacy and self-esteem issues. When faced with new challenges such as learning a new form of technology, there is a high probability that these individuals would face some problems or simply try to avoid the new situation altogether.

Conscientious users show a positive correlation with networksize (0.14) but a negative correlation with density (−0.14), as their friend network is sparse and large, and their friends are dispersed socially. Users high in agreeableness are often selected as friends, and they themselves tend to choose friends with similar agreeableness, extraversion, and openness scores [56].

Brokerage and betweenness is the extent to which an individual is connected to people or clusters of people who are not connected to each other as a broker [37], [38]. Our results revealed that among the Big 5 personality traits, the highest correlation results between brokerage and betweenness occurred for extraverted individuals (0.25). Conscientiousness users are considered as hard working and organized individuals. They might also be selected into brokerage (0.11) and betweenness roles (0.11), especially in instrumental networks and as work partners. They are chosen especially when colleagues from different organizational areas seek somebody out for resolving work-related problems. Positive results on brokerage and betweenness are also beneficial for users high in agreeableness (0.05). They usually show empathy, tend to be cooperative and motivated to develop positive relations with others. These characteristics make them attractive to friendship relationships as they are often chosen as friends over time and invited into the team friendship networks [56]. They are more likely to relate positively to a number of ties directed to the node and help integrate conflicting partners' views and needs. Users high in openness, as creative and open-minded people, are also positively correlated with brokerage (0.04) and betweenness (0.04), hence, they exhibit diverse interests. Due to their curiosity, they are interesting as conversational partners and are more likely to be chosen for friendships. In their pursuit of ties to contacts from different, unconnected social circles, they might serve as network brokers.

In contrast to other personality types, neurotics, as anxious, insecure and hostile individuals, correlate negatively with brokerage (−0.13) and betweenness (−0.03). In other words, they relate negatively to any team friendship and advice networks [57]. They often express negative emotions and may be viewed as high-cost interaction partners who are likely to be avoided. This finding might explain the reason why most neurotic users tend to update their statuses with anger words (0.20), use words indicating positive emotions (−0.14) and social interaction (−0.13) as shown in Table 4 and Table 5.

VI. PREDICTION

After the analysis of the significant correlations determined in the previous section, we conducted an experimental evaluation based on three baseline methods and the XGBoost model as the primary classifier. The aim of the evaluation was to investigate the predictive ability of the models using the myPersonality dataset containing the entire 96-dimension feature space for both individual and combined features. First, we evaluated each method with single feature sets, that is SNA, LIWC and SPLICE features, before we tested the predictive power of the models based on the combined sets. Personality prediction results presented in the next section were obtained by using the XGboost algorithm as the primary classifier. We trained our data with an open project implemented in Python, a popular machine learning workbench, 10-fold cross-validation with 10 iterations. Each time, a single fold was used for testing, and the other 9 folds were used for training. For comparison, we chose three machine learners, specifically, Support Vector Machine (SVM), Logistic Regression and Gradient Boosting as a baseline for comparison with the primary classifier.

According to the experiment results presented in Table 7, we can see that the XGBoost approach outperformed the average baseline for all the feature sets. It was significantly higher in accuracy for all the personality traits and confirmed the theory of XGBoost as an efficiently fast and scalable machine learning system [58]. In comparison to other gradient boosting machines, XGBoost uses a more regularized-model formalization to control over-fitting and provides a better performance [59]. With the combination of all the features we extracted, we can predict personality traits with an average accuracy of 74.2%.

Based on the previous study, features belonging to individual categories are often combined to maximize the model's performance [60]. The results of our study, however, show that the highest performance with the XGBoost algorithm was achieved for extraversion by using the individual SNA features sets rather than the combined sets. The prediction accuracy for extraversion reached 78.6% in comparison to the combined set with the accuracy of 74.2%. In addition, the performance accuracy for other traits such as openness, neuroticism and consciousness in individual feature sets was also higher than that for the combined features. Agreeableness, however, was the only trait with lower prediction accuracy scores for the individual features than for the combined

TABLE 7. Prediction result table.

		O	N	A	C	E
Feature	Algorithm	acc. %	acc. %	acc. %	acc. %	acc. %
SNA	XGB	73.3	68.0	65.3	69.8	78.6
	LR	70.0	57.60	52.1	53.6	51.6
	GB	61.2	58.8	50.8	48.4	68.4
	SVM	58.8	38.8	41.6	53.6	68.0
LIWC	XGB	73.3	66.1	64.1	61.3	60.2
	LR	70.4	60.4	53.6	53.6	61.6
	GB	62.3	58.8	56.4	54.4	60.8
	SVM	70.4	60.4	52.4	56	61.6
SPLICE	XGB	71.9	62.6	64	61.3	59.7
	LR	65.6	57.6	52	52	56.4
	GB	63.6	63.6	51.2	50.8	54.0
	SVM	44.2	54.8	51.2	48.4	51.6
SNA+LIWC+ SPLICE	XGB	73.1	63.2	59.0	62.4	74.2
	LR	68.0	61.6	44.0	54.8	64.8
	GB	61.6	52.0	49.6	49.6	60.4
	SVM	62.8	46.4	56.8	50.8	63.2

[a]acc. represents accuracy; LR = Logistic Regression, GB = Gradient Boosting; XGB = XGBoost, SVM = Support Vector Machine.

sets, that is 65.3% for SNA, 61.3% for SPLICE and 61.3% for LIWC. The accuracy for the combined feature set was 62.4%.

The analysis of the performance accuracy for, individual SNA and linguistic features sets showed that using SNA features that infer the user's activity behavior for four personality traits showed overwhelming results. Contrary to linguistic features, our results reveal that even though SNA was trained with a small number of correlated features, it still achieved a higher performance accuracy. Specifically, the performance accuracy was 78.6% for extraversion, 73.3% for openness, 68% for neuroticism and 69.8% for the conscientiousness trait. Only agreeableness, however, with 65.3% for the SNA features, had a lower performance accuracy than that for the linguistic feature sets reaching 61.3%. Our results clearly illustrate a higher potential for individual SNA features for personality prediction accuracy based on user behavior on social networks.

The prediction accuracy findings of the linguistic feature sets indicate that using LIWC features in personality prediction can result in a higher performance than that with SPLICE. Our study shows that the highest prediction performance was reached for openness and neuroticism with 73.3% and 66.1% accuracy, respectively. The accuracy of these traits were followed by agreeableness and conscientiousness with 64.1% and 61.3%, respectively. Extraversion had a lower prediction performance with SPLICE reaching 59.7% and with LIWC features at 60.2%.

The results for prediction accuracy show that even if tested under the same dataset, the personality prediction system built on the XGBoost algorithm performed significantly better than those built on logistic regression, the gradient boosting classifier or the support vector machine. The future development of our study may utilize a larger training dataset which will allow the system to include a wider variety of features to increase the system's accuracy.

VII. CONCLUSION AND FUTURE WORK

In this paper, we provide an outline of insights for research on social networks and personality psychology. The study investigates the literature on the uses of social media framework as behavioral feature study by exploring the relationship between users' personalities and their behaviours in social networks. To predict a user's personality, we conducted a comparative study of best behavioral indicators for Facebook usage of the same set of features to capture the ways the users socialize, communicate and connect with each other. To perform our research, we used myPersonality dataset to design a large set of features that play an important role in determining different personality traits. Our results show that a great amount of insight can be gained from studying the social and linguistic indicators of personality. We found that using different linguistic dictionaries can be helpful in improving the correlation results. We realized that the linguistic features, due to their large numbers, rich different correlation varieties; in comparison to social network features, they make intuitive sense. Computing the Pearson correlation values between the dataset and each personality dimensionality showed that different personalities match different types of features. In the next section of our study, we used four machine learning algorithms to predict the personality scores from extracted features. Our results showed that using social network features for personality prediction can achieve a higher performance than using the linguistic features. The highest personality prediction was performed with the XGBoost machine learning approach by using individual SNA features sets. Specifically, the highest accuracy was achieved for extraversion as the trait that was most often expressed by Facebook features. The performance accuracy for openness, neuroticism and agreeableness in individual feature sets was also higher than in combined features. Overall, with the combination of all the features we extracted, we predicted the personality traits with an average accuracy of 74.2%. These results illustrate a higher potential of individual social network features for personality prediction. Inferring the personality traits of users in Facebook social media platform not only helped us understand the users' online behavior, but also offers us a guidance for personalized services improvements in the future. For future work, there are several important areas to improve our scope of research. Since our experiment was based on a small number of items from the myPersonality sample dataset (250 users, 9917 status updates), the accuracy of the results tended to be rather limited. We need to utilize a larger training dataset which will allow the system to include itself in a wider variety of feature sets to increase the systems accuracy. With this improvement we will be able to answer more practical questions, such as how to recommend socially relevant and well-presented information to users based on the mutuality between the nodes in their social network groups. Examining personalities from Facebook profile statuses may allow recommender systems to improve their prediction accuracy by recommending items, such as TV shows, music or sports events designed in accordance to the user's personality. The items could be recommended to an individual user based on the ratings of mutual connections. Moreover, by using a collaborative filtering technique, we could select users with similar tastes and recommend the items to them. This could be applied to the individuals who share similar personality traits. Developing and evaluating these approaches is a new space open for future work.

ACKNOWLEDGMENT

The authors would like to thank Ms. Janka Koperdanova for her full support and editing of the paper, and Mr. Mussie Afeworki and Mr. Tesfay Gebremariam for their technical help.

REFERENCES

[1] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland, "Friends don't lie: Inferring personality traits from social network structure," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 321–330.

[2] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Tech. Rep., 2015.

[3] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, "Lexical predictors of personality type," Tech. Rep., 2005.

[4] R. R. McCrae and P. T. Costa, "Personality, coping, and coping effectiveness in an adult sample," *J. Pers.*, vol. 54, no. 2, pp. 385–404, 1986.

[5] B. R. Karney and T. N. Bradbury, "The longitudinal course of marital quality and stability: A review of theory, method, and research," *Psychol. Bull.*, vol. 118, no. 1, pp. 3–34, 1995.

[6] P. T. Costa and R. R. McCrae, "Normal personality assessment in clinical practice: The NEO personality inventory," *Psychol. Assessment*, vol. 4, no. 1, pp. 5–13, 1992.

[7] S. Adali and J. Golbeck, "Predicting personality with social behavior," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2012, pp. 302–309.

[8] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of Personality: Theory and Research*, vol. 2. 1999, pp. 102–138.

[9] S. D. Gosling, A. A. Augustine, S. Vazire, N. Holtzman, and S. Gaddis, "Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information," *Cyberpsychol., Behav., Social Netw.*, vol. 14, no. 9, pp. 483–488, 2011.

[10] H. Davis *et al.*, in *Proc. 3rd ACM Web Sci. Conf. (WebSci)*, Paris, France, 2013.

[11] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *Proc. Extended Abstr. Hum. Factors Comput. Syst. (CHI)*, 2011, pp. 253–262.

[12] C. Ross, E. S. Orr, M. Sisic, J. M. Arseneault, M. G. Simmering, and R. R. Orr, "Personality and motivations associated with Facebook use," *Comput. Hum. Behav.*, vol. 25, no. 2, pp. 578–586, 2009.

[13] C. Sumner, A. Byers, and M. Shearing, "Determining personality traits & privacy concerns from Facebook activity," *Black Hat Briefings*, vol. 11, pp. 197–221, Dec. 2011.

[14] Y. Kalish and G. Robins, "Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure," *Social Netw.*, vol. 28, no. 1, pp. 56–84, 2006.

[15] I. F. Iatan, "Predicting human personality from social media using a fuzzy neural network," in *Issues in the Use of Neural Networks in Information Retrieval*. Springer, 2017, pp. 81–105.

[16] H. A. Schwartz *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, p. e73791, 2013.

[17] T. Yarkoni, "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers," *J. Res. Pers.*, vol. 44, no. 3, pp. 363–373, 2010.

[18] M. Kosinski, S. C. Matz, S. D. Gosling, V. Popov, and D. Stillwell, "Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines," *Amer. Psychol.*, vol. 70, no. 6, pp. 543–556, 2015.

[19] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell, "Personality and patterns of Facebook usage," in *Proc. 4th Annu. ACM Web Sci. Conf.*, 2012, pp. 24–32.

[20] G. Farnadi, S. Zoghbi, M.-F. Moens, and M. De Cock, "Recognising personality traits using Facebook status updates," in *Proc. Workshop Comput. Pers. Recognit. (WCPR13), 7th Int. AAAI Conf. Weblogs Social Media (ICWSM)*, 2013.

[21] I. Cantador, I. Fernández-Tobías, and A. Bellogín, "Relating personality types with user preferences in multiple entertainment domains," in *Proc. CEUR Workshop Shlomo Berkovsky*, 2013.

[22] T. Tandera *et al.*, "Personality prediction system from Facebook users," *Procedia Comput. Sci.*, vol. 116, pp. 604–611, Dec. 2017.

[23] V. Ong *et al.*, "Personality prediction based on twitter information in Bahasa Indonesia," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, 2017, pp. 367–372.

[24] L. R. Goldberg, "The structure of phenotypic personality traits," *Amer. Psychol.*, vol. 48, no. 1, pp. 26–34, 1993.

[25] E. C. Tupes and R. E. Christal, "Recurrent personality factors based on trait ratings," *J. Pers.*, vol. 60, no. 2, pp. 225–251, 1992.

[26] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference," *J. Pers. Social Psychol.*, vol. 77, no. 6, pp. 1296–1312, 1999.

[27] T. Morton, J. Kottmann, J. Baldridge, and G. Bierner, "Opennlp: A java-based NLP toolkit," in *Proc. EACL*, 2005.

[28] V. Evrim and A. Awwal, "Effect of personality traits on classification of political orientation," *World Acad. Sci., Eng. Technol., Int. J. Social, Behav., Educ., Econ., Bus. Ind. Eng.*, vol. 9, no. 6, pp. 2001–2006, 2015.

[29] T. P. Michalak, T. Rahwan, and M. Wooldridge, "Strategic social network analysis," in *Proc. AAAI*, 2017, pp. 4841–4845.

[30] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, 2010.

[31] K. Moffitt, J. Giboney, E. Ehrhardt, J. Burgoon, and J. Nunamaker, "Structured programming for linguistic cue extraction (SPLICE)," in *Proc. HICSS Rapid Screening Technol., Deception Detection Credibility Assessment Symp.*, 2012, pp. 103–108.

[32] A. J. Omalley and P. V. Marsden, "The analysis of social networks," *Health Services Outcomes Res. Methodol.*, vol. 8, no. 4, pp. 222–269, 2008.

[33] D. D. Greene, "Graph and network analysis?" Tech. Rep., 2011.

[34] J. H. Fowler and N. A. Christakis, "Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study," *Brit. Med. J.*, vol. 337, p. a2338, Sep. 2008.

[35] J.-E. Lönnqvist, J. V. Itkonen, M. Verkasalo, and P. Poutvaara, "The five-factor model of personality and degree and transitivity of Facebook social networks," *J. Res. Person.*, vol. 50, pp. 98–101, Jun. 2014.

[36] D. J. Brass, "Being in the right place: A structural analysis of individual influence in an organization," *Admin. Sci. Quart.*, vol. 29, no. 4, pp. 518–539, 1984.

[37] S. E. Seibert, M. L. Kraimer, and R. C. Liden, "A social capital theory of career success," *Acad. Manage. J.*, vol. 44, no. 2, pp. 219–237, 2001.

[38] R. S. Burt, *Brokerage and Closure: An Introduction to Social Capital*. London, U.K.: Oxford Univ. Press, 2005.

[39] H. Lin and L. Qiu, "Sharing emotion on Facebook: Network size, density, and individual motivation," in *Proc. Extended Abstr. Hum. Factors Comput. Syst. (CHI)*, 2012, pp. 2573–2578.

[40] L. C. Freeman, D. Roeder, and R. R. Mulholland, "Centrality in social networks: II. Experimental results," *Social Netw.*, vol. 2, no. 2, pp. 119–141, 1979.

[41] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 68, no. 3, p. 036122, 2003.

[42] M. Aghagolzadeh, I. Barjasteh, and H. Radha, "Transitivity matrix of social network graphs," in *Proc. IEEE Stat. Signal Process. Workshop (SSP)*, Aug. 2012, pp. 145–148.

[43] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 856–863.

[44] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Using Twitter content to predict psychopathy," in *Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA)*, vol. 2, Dec. 2012, pp. 394–401.

[45] D. Cramer, *Fundamental Statistics for Social Research: Step-by-Step Calculations and Computer Techniques Using SPSS for Windows*. Evanston, IL, USA: Routledge, 2003.

[46] M. A. Hall, *Correlation-Based Feature Selection for Machine Learning*. 1999.

[47] G. Farnadi *et al.*, "How are you doing?: Emotions and personality in Facebook," in *Proc. EMPIRE (2nd Workshop Emotions Personality Personalized Services), Workshop UMAP (22nd Conf. User Modeling, Adaptation Personalization)*, 2014, pp. 45–56.

[48] S. Bai, R. Gao, and T. Zhu, "Determining personality traits from Renren status usage behavior," in *Computational Visual Media*. Springer, 2012, pp. 226–233.

[49] H. Wei *et al.*, "Beyond the words: Predicting user personality from heterogeneous information," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, 2017, pp. 305–314.

[50] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," *Comput. Hum. Behav.*, vol. 26, no. 6, pp. 1289–1295, 2010.

[51] Y. A. Hamburger and E. Ben-Artzi, "The relationship between extraversion and neuroticism and the different uses of the Internet," *Comput. Human Behavior*, vol. 16, no. 4, pp. 441–449, 2000.

[52] J. M. Balmaceda, S. Schiaffino, and D. Godoy, "How do personality traits affect communication among users in online social networks?" *Online Inf. Rev.*, vol. 38, no. 1, pp. 136–153, 2014.

[53] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: A meta-analysis," *Pers. Psychol.*, vol. 44, no. 1, pp. 1–26, 1991.

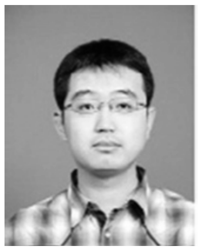
[54] G. M. De Jong, E. van Sonderen, and P. M. Emmelkamp, "A comprehensive model of stress," *Psychotherapy Psychosomatics*, vol. 68, no. 6, pp. 290–298, 1999.

[55] L. R. Goldberg, "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models," *Person. Psychol. Eur.*, vol. 7, no. 1, pp. 7–28, 1999.
[56] M. Selfhout, W. Burk, S. Branje, J. Denissen, M. Van Aken, and W. Meeus, "Emerging late adolescent friendship networks and big five personality traits: A social network approach," *J. Pers.*, vol. 78, no. 2, pp. 509–538, 2010.
[57] K. J. Klein, B.-C. Lim, J. L. Saltz, and D. M. Mayer, "How do they get there? An examination of the antecedents of centrality in team networks," *Acad. Manage. J.*, vol. 47, no. 6, pp. 952–963, 2004.
[58] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
[59] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[60] Y. Wang, "Understanding personality through social media," Tech. Rep.

MICHAEL M. TADESSE received the B.S. degree in management information system from the Unity University College, Addis Ababa, Ethiopia, in 2005, and the M.S. degree in software engineering from the Chongqing University of Technology in 2011, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Dalian University of Technology. His current research interests include personality prediction, recommendation system information retrieval, and natural language processing.

HONGFEI LIN received the B.Sc. degree from Northeastern Normal University in 1983, the M.Sc. degree from the Dalian University of Technology in 1992, and the Ph.D. degree from Northeastern University in 2000. He is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology. He is also the Director of the Information Retrieval Laboratory, Dalian University of Technology. He has published over 100 research papers in various journals, conferences, and books. His research projects are funded by the National Natural Science Foundation of China and the National High-Tech Development Plan. His research interests include information retrieval, text mining, natural language processing, and effective computing. In recent years, he has focused on text mining for biomedical literatures, biomedical hypothesis generation, information extraction from huge biomedical resources, learning to rank, sentimental analysis, and opinion mining.

BO XU received the B.Sc. degree from the Dalian University of Technology, China, in 2011, where he is currently the Ph.D. degree with the School of Computer Science and Technology. His current research interests include information retrieval, learning to rank, natural language processing, and biomedical text mining.



LIANG YANG received the B.Sc. degree from the Dalian University of Technology, China, in 2009, and the Ph.D. degree in 2017. He is currently a Lecturer with the School of Computer Science and Technology, Dalian University of Technology. His current research interests include sentiment analysis, natural language processing, and text mining.