Comparative Analysis of Feature Selection Algorithms for Computational Personality Prediction From Social Media

Ahmed Al Marouf, Md. Kamrul Hasan, and Hasan Mahmud

***Abstract*— With the rapid growth of social media, users are getting involved in virtual socialism, generating a huge volume of textual and image contents. Considering the contents such as status updates/tweets and shared posts/retweets, liking other posts is reflecting the online behavior of the users. Predicting personality of a user from these digital footprints has become a computationally challenging problem. In a profile-based approach, utilizing the user-generated textual contents could be useful to reflect the personality in social media. Using huge number of features of different categories, such as traditional linguistic features (character-level, word-level, structural, and so on), psycholinguistic features (emotional affects, perceptions, social relationships, and so on) or social network features (network size, betweenness, and so on) could be useful to predict personality traits from social media. According to a widely popular personality model, namely, big-five-factor model (BFFM), the five factors are openness-to-experience, conscientiousness, extraversion, agreeableness, and neuroticism. Predicting personality is redefined as predicting each of these traits separately from the extracted features. Traditionally, it takes huge number of features to get better accuracy on any prediction task although applying feature selection algorithms may improve the performance of the model. In this article, we have compared the performance of five feature selection algorithms, namely the Pearson correlation coefficient (PCC), correlation-based feature subset (CFS), information gain (IG), symmetric uncertainly (SU) evaluator, and chi-squared (CHI) method. The performance is evaluated using the classic metrics, namely, precision, recall, f-measure, and accuracy as evaluation matrices.**

***Index Terms*— Chi-squared (CHI) method, computational personality prediction, feature selection algorithms, information gain (IG), Pearson correlation coefficient (PCC), social media.**

I. Introduction

SOCIAL media platforms such as Facebook, Twitter, Google+, and Instagram has gained popularity due to ease access throughout the world and user-friendly interfaces to start communicating with others within a short period of time. Each user in these social networking sites (SNSs) is considered as an entity, and each entity is connected with other entities as friends, connections, or followers. While using these SNS's, users are facilitated by many activities, such as posting statuses/tweets, sharing others' posts/retweets, liking others' posts, commenting on others' posts, chatting directly with the friends, and playing online games with the friends. It is evident that from the activities performed by users, online behavior could be depicted [1]. Understanding users' behavior may help to identify personality traits.

Predicting users' personalities from digital footprints of social media is a challenging task as the context of identifying personality traits in social media is not trivial. Users behave differently in social media and real life. Therefore, the user-generated content, such as status updates in social media, may provide enough evidential reflection of personality as SNS user posts statuses based on his/her current situation, a recent political or popular event, hyped topics, and so on. For example, during an election of his/her country, he/she may posts positive or negative reviews/opinions about a political party. These types of statuses may have contextual trends, as other friends of the users may also be involved in posting similar statuses. Considering the trend, user may post his/her political views. Users are creating trends as well as following different trends to become popular or socially accepted by their friends in social media. Moreover, each user has different perceptions and different interest categories to be triggered to update statuses. For defining personality, we have followed the widely used big-five-factor model (BFFM). According to BFFM, there are four positive personality traits, namely, openness-to-experience (O), conscientiousness (C), extraversion (E), agreeableness (A), and the only negative trait neuroticism (N). This personality model is also known as the OCEAN model.

It is evident that user-generated content could be an effective data source to build a predictive model [2]. The status updates posted by the SNS users have the influence of culture and personal issues. The structures of various languages actually influence on identity, culture, and diversity of persons [3]. Therefore, Facebook status has become a research tool for the researchers for identifying the personality [4]. Therefore, a model could be built based on supervised learning systems to predict personality traits from Facebook statuses.

Feature extraction and feature selection are applied afterward to identify the most relevant features. Those features are trained to a classification model, and testing is performed afterward. Therefore, finding the most relevant feature is one of the challenging tasks to be performed to get better accuracy

Manuscript received April 13, 2019; revised August 31, 2019 and October 31, 2019; accepted December 15, 2019. *(Corresponding author: Ahmed Al Marouf.)*

Ahmed Al Marouf is with the Department of Computer Science and Engineering, Daffodil International University (DIU), Dhaka 1207, Bangladesh (e-mail: ahmedalmarouf@gmail.com).

Md. Kamrul Hasan and Hasan Mahmud are with the Department of Computer Science and Engineering, Islamic University of Technology (IUT), Gazipur 1704, Bangladesh (e-mail: hasank@iut-dhaka.edu; hasan@iut-dhaka.edu).

Digital Object Identifier 10.1109/TCSS.2020.2966910

over the prediction. One of the main contributions of this article is to identify the best feature selection algorithm for extracting the most prominent features among the traditional linguistic, psycholinguistic, and social network (SN) features. As the number of features is relevantly high and the accuracy is low while applying all the features, we have investigated several cases to find the important features and category of features for predicting personality precisely.

In this article, we try to compare the existing feature selection algorithms to predict personality from Facebook status updates. In summary, we have the following contributions.

1) Applying different feature selection algorithms, such as the chi-squared (CHI) method, Pearson coefficient correlation, information gain (IG), correlation-based feature subset (CFS)-based subset evaluation, and symmetrical uncertainty attribute evaluation, to predict the big-five-personality traits.
2) We have extracted over 150 features to analyze the predictive system over different types of features, such as traditional linguistic, psycholinguistic, and SN features. In the literature, many researchers have used few features to predict personality, but the overall outcome of those approaches is not quite satisfactory. Jelling up a huge volume of features has given an evidentially better understanding of personality traits.
3) We have considered several scenarios/cases of feature combinations based on psycholinguistic features to find the best subset of features to predict each personality trait differently. Hence, we have determined the accuracy with and without SN features that are reported in the experiments.
4) Five different classifiers, namely, the naïve Bayes (NB), decision tree (DT), random forest (RF), simple logistic regression (SLR), and support vector machine (SVM), were used to determine the evaluation metrics to find the best feature selection algorithm. Utilizing these classifiers, we derived several conclusions.

II. Related Works

In this section, we have discussed state-of-the-art works regarding predicting personality traits and applying feature selection algorithms. This section is divided into several parts: literature of computational personality prediction, literature of the psycholinguistic tools, literature of existing methods and applied features devised by researchers, and feature selection methods applied in similar research problems.

For predicting personality traits computationally, researchers have utilized the machine learning techniques, such as supervised/ unsupervised learning models and classification algorithms to classify the traits. International personality item pools (IPIPs) [5] are the items or questions to answer to devise a scoring mechanism for traits identification. Depending on the behavior of test-taker on different issues of practical life, these items are presented. Using the IPIP questionnaire, the quantitative method has been adopted for the problem, and many variations of the question sets were used for developing a better ground-truth data set. This manual procedure of taking answers of a set of questionnaire could be easily adopted. However, the main limitation of this process is the test-takers need to answer the questions honestly. The different IPIP sets are discussed in Section III, elaborately.

The correlation between the usage of Facebook, thus, social media, and personality has been studied in [6] and [7]. In [6], the study shows that the correlation is higher for neuroticism and extraversion trait but average for the other traits. Different literature works establish the relationships between the personality and social media uses, such as personality of popular social media users [8], influence of personality from Facebook usage, and wall posting [9], by mining social interactions in Facebook [10], capturing personality from photograph or photograph-related posts in social media [11], and so on.

Howlader *et al.* [12] proposed the topic modeling-based approach applied to Facebook status updates. For this work, they used the linear Dirichlet allocation (LDA) and term frequency-inverse document frequency (TF-IDF) as feature and applied flexible regression models for prediction. Deep learning-based methods were introduced by Tandera *et al.* [13]. They applied the traditional deep learning algorithms, such as the multilayer perceptron (MLP), long short-term memory (LSTM), gated recurrent unit (GRU), and 1-D convolutional neural network (CNN-1-D). A huge feature set (725 features) has been analyzed in [14] considering basic linguistic features, POS-tagger parameters, AFINN (Lexicon list) parameters, and H4Lvd parameters. A review of emerging trends of personality prediction from online social media is performed by Kaushal and Patwardhan [15]. They listed different categories of features, such as linguistic features (LIWC features, POS tags, speech acts, and sentiment features), nonlinguistic features (structural, behavior, and temporal features), and SN features. Based on the features used for identifying personality traits, the methodologies have modified. Farnandi *et al.* [46] have proposed methods for predicting personality from social media considering the cross-platform and cross-domain situations. Considering personality prediction as a multilabel prediction task, they have extracted the LIWC, MRC, NRC, and SPLICE features to run several types of regression models.

For extracting relevant psychological features from texts, psycholinguistic tools are utilized. These software tools are developed for easier experimentations. LIWC [16], MRC [17], and SPLICE [18] are widely used psycholinguistic tools. Developed by Pennebaker and Francis, a word list-based text analysis tool, LIWC, extracts 93 features consisting standard counts (word counts, words longer than six letters, and so on), personal concerns (occupation, financial issues, health, and so on), psychological processes (cognitive, emotional, perceptional, and social processes), and other features (punctuation counts, swear words, and so on) [16]. On the other hand, MRC [17] features are computed using Medical Research Council's psycholinguistic database that consists over 150 000 words with linguistic and psycholinguistic features of each word. MRC includes very interesting latent features of text, such as the Kucera–Francis written frequency [19] and the Brown verbal frequency [20]. Structured Programming for Linguistic Cue Extraction (SPLICE) extracts 74 features related to linguistic. Upon the

input of textual data, SPLICE [18] outcomes various features including the quantities (number of characters, sentences, words, and so on), parts of speech features (number of nouns, noun ratio, verb ratio, adjective ratio, and so on), immediacy (number of passive verbs and passive verb ratio), pronouns, positive self-evaluation, negative self-evaluation, influence, deference, and whissel (imagery, pleasantness, and activation), text complexity, spoken word features, tense, SentiWordNet features, and readability scores. Among these three widely used closed vocabulary psycholinguistic tools, for our work, we have used LIWC. LIWC consists of a psycholinguistic dictionary in back end, which contains huge number of words, synonyms, and antonyms in different psychological categories. LIWC is proven to be useful in the context of personality traits prediction.

Though features are playing a vital role in data-driven system, the feature selection methods also significantly find the most prominent features from huge feature vectors. Not only, specifically, data mining but also feature selection has become an important tool used in bioinformatics and computational biology. Xu *et al.* [51] proposed an autoencoder-based feature selection method for classification of anticancer drug response. Similarly, Mallik and Zhao [52] presented a graph- and rule-based learning algorithm for cancer-type classification using feature selection. Mallik and Zhao [53] have applied statistically significant feature extraction-based study on cancer expression using integrated marker recognition, which is mutual-information based.

Apart from the filter-based feature selection algorithm, there are wrapper based and hybrid feature selection algorithms. Masoudi-Sobhanzadeh *et al.* [57] have presented "FeatureSelect," which is a software for selecting features based on machine learning approaches, and the software is tested on gene selection methods. Several nature-inspired evolutionary algorithm-based feature selection algorithms are presented recently. Mafarja *et al*. [58] presented a binary grasshopper optimization algorithm-based feature selection. Similarly, chaotic hybrid artificial bee colony-based feature selection [59] and binary butterfly optimization-based feature selection [60] are recently introduced in the literature.

In the process of supervised learning, one of the most significant roles is played by the feature selection criteria. Selecting the most relevant features from a huge feature vector has a vital impact on the accuracy of the system. For comparing, in this article, we have utilized the five most conventional feature selection algorithms, namely IG, CFS-based subset evaluator (CFS), CHI method, symmetrical uncertainty attribute evaluation (SU), and the Pearson correlation coefficient (PCC). These feature selection algorithms are discussed in Section III, including the definitions and formulas.

Therefore, in this article, we have presented an experimental comparison between the feature selection algorithms, and for the experiments, we have extracted more than 150 features. The rest of this article is organized as follows. Section III discusses the computational personality-prediction problem, and Section IV includes the state-of-the-art application areas of the feature selection algorithms. Section V illustrated the proposed experimental method and experimental results are enlisted in Section VI. The detail comparative analysis is depicted in Section VII, and finally, Section VIII concludes with the contributions highlighted.

III. Computational Personality-Prediction Problem

The computational personality-prediction problem in the context of social media could be defined as "predicting the personality traits from user profile information using computational features rather than asking a set of questionnaire." Usually, for understanding own personality, people try to take online or off-line personality test. The traditional personality-prediction systems depend on a set of questionnaires to be answered honestly by the test-taker. Questionnaire-based personality-prediction systems are also popular among the test-takers. The widely used personality tests are big-five-personality test [20], the Myers–Briggs type indicator (MBTI) [21], and the dominance influence steadiness conscientiousness (DISK) [22]. Among these tests, the big-five-personality test has been widely accepted among the test-takers because of the similarity found with themselves with the result of the test.

Many online personality testing sites, such as 16Personality[1], 123test[2], Personality Perfect[3], PsychCentral Personality Test[4], Open Source Psychometrics Project[5], See My Personality[6], and Discover My Profile[7] by the University of Cambridge, are very popular for identifying precise personality reviewed by the test-takers. The reviews are analyzed from each of the websites and found positive comments delivered by the reviewers. The literature provides evidential proof that computational personality prediction provides better results than manual paper-based methods. Therefore, the acceptability of these online personality tools is much higher than manual questionnaire-based personality testing. Hence, this encourages applying automated personality prediction from social media. It is evident that computational personality judgments are more accurate than those made by humans [32].

The history of personality prediction goes a long way as researchers have tried to optimize the number of questions being asked to the test-taker. Usually, high volume of questions is asked, and the answers are analyzed to predict personality precisely. However, answering these questions could be time-consuming as well as tiring for the test-takers. Therefore, asking a minimum number of questions to get a better prediction could be a challenging task. Researchers' have come up with various numbers of questions or items. NEO five-factor inventory (NEO-FFI) [24] is a 60-item personality measure model. Similar models were proposed by researchers in psychology area for the personality-prediction task. Depending on scores determined by the IPIP, the computation of

[1]https://www.16personalities.com/
[2]https://www.123test.com/
[3]https://www.personalityperfect.com/
[4]https://psychcentral.com/personality-test/
[5]https://openpsychometrics.org/
[6]http://www.seemypersonality.com/
[7]https://discovermyprofile.com/

personality traits is performed. Depending on the number of IPIP items considered for prediction, there are several models proposed by many researchers. The 50-item IPIP five-factor model (FFM) proposed by Goldberg [25], 44-item big-five inventory (BFI) proposed by John and Srivastava [26], 40-item Big-Five Mini-Markers proposed by Saucier [27], 20-item Mini-IPIP proposed by Donnellan *et al.* [28], ten-item personality inventory (TIPI) proposed by Gosling *et al.* [29] are the existing models in the literature. Short form of item sets is also proven effective in some cases [30]. Although there are many scoring systems adopted for this particular problem, each of them has own advantages to be used. The myPersonality data set [4], [31] is collected from Facebook users and used 100-item IPIP questionnaire set.

Although the above-mentioned works try to use psycholinguistic tools to extract the psycholinguistic features or different types of methods for devising a model, in the literature, there is no article highlighting the best features for predicting personality from social media data. In this article, we have focused on this problem and designed experiments to find a solution to this problem.

IV. Feature Selection Algorithms Overview

In this section, we have outlined the feature selection algorithms that we have applied for computational personality prediction. The five different algorithms are applied to the outcome of features that are most relevant to the prediction task. All these feature selection methods provide a ranking generated based on the relevance between the feature and the class.

CFS subset evaluator [33], [34] is a feature selection algorithm that finds the subset of features via the individual predictive ability of each feature along with the degree of redundancy between them. CFS ranks the features subsets according to a correlation-based heuristic evaluation method.

The subset evaluation function is given in (1), where *Ms* is the heuristic merit of the feature subset S containing k features. $\bar{r_{cf}}$ is the mean of feature-class correlation ($f \in S$), and $\bar{r_{ff}}$ is the average feature-feature intercorrelation

$$M_s = \frac{k\bar{r_{cf}}}{\sqrt{k + k(k-1)\bar{r_{ff}}}}. \tag{1}$$

The CFS subset evaluator is used in different contexts of research, such as to predict students' performance [35] and selecting features for sentiment classification [36].

IG is one of the widely used feature selection methods in different research problems, including text categorization [37], [38]. Various research fields have utilized the inner mechanism of IG, such as computer vision and text classification [38], [39]. IG outcomes a ratio value calculated by (2), where values(a) denotes the set of all possible values of features $a \in$ Attr. Attr is the set of all features, H is the entropy, and $x \in T$ denotes the value of specific example x for $a \in$ Attr. The largest IG is the smallest entropy

$$\mathrm{IG}(T, a) = H(T) - \sum_{v \in \mathrm{vals}(a)} \frac{|\{x \in T \mid x_a = v\}|}{|T|} \cdot H(x \in T | x_a = v). \tag{2}$$

In the context of statistics, the uncertainty coefficient or entropy coefficient is the measure of nominal association. The symmetrical uncertainty (SU) [39] attribute evaluator is one kind of correlation finder that evaluates the importance of a feature by measuring the SU with respect to the class. This feature selection process is not only used for imagery data, such as hyperspectral images [40], but also used with the nature-inspired optimization algorithms, such as ant colony optimization [41]. The SU is determined using (3) where, *H(C|F)* is the conditional entropy considering C the class and F the feature, and $H(C)$ is the single distribution of class C. The algorithm outcomes ranking of the most relevant features

$$\mathrm{SU}(C, F) = \frac{2 * (H(C) - H(C|F))}{H(C) + H(F)} \tag{3}$$

$$H(C) = -\sum_{x} P_C(x) \log P_C(x) \tag{4}$$

$$H(C \mid F) = -\sum_{x,y} P_{C,F}(x, y) \log P_{C,F}(x, y). \tag{5}$$

CHI test (χ^2) [42], [43] is used in statistics for determining the association between variables or features. Depending on the difference between the expected frequencies (e) and the observed frequencies (n) in one or more features in the feature set, the CHI value is determined. Depending on the value of the parameter, we can decide the number of features to be selected for a system. The equation for calculating CHI value is given as

$$\chi 2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{6}$$

where r and c are the numbers of row and column of the feature table.

PCC [44] is considered as one of the most efficient and widely accepted feature selection algorithm. In PCC, the value of covariance between the class and feature is been determined. The standard deviations (SDs) of the class and feature are calculated to find the coefficient value (ρ). The coefficient could be used as an efficient parameter to determine the feature sets. The calculation of (ρ) is performed using (7) as given in the following. cov*(X, Y)* is the covariance between X and Y, where X or Y is the class value, and σ is the SD in the following equation:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{7}$$

The correlation value is distributed between -1 and $+1$, where 1 is the total positive correlation, O is no linear correlation, and -1 means the negative correlation. The coefficient is invariant under separate modifications in scale and location in the two variables, which could be considered as a key mathematical property of PCC. Depending on the above-mentioned parameters, the number of features to be selected for the problem could be determined. For comparing the feature selection criteria, we have experimented with various scenarios for computational personality trait prediction.

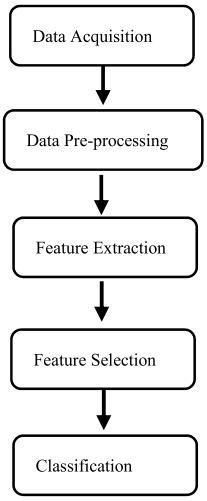

Fig. 1. Steps of the experimental method.

TABLE I
CLASS DISTRIBUTION OF MYPERSONALITY DATA SET

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

V. EXPERIMENTAL METHOD

For the experimental analysis, we have designed a common method for testing the performance of each of the feature selection algorithms. The proposed experimental method consists of data acquisition, data preprocessing, feature extraction, feature selection, and classification, as depicted in Fig. 1. For each of the five personality traits, we are going to apply the proposed method. The rest of the section elaborately discusses the steps of experimental methods.

A. Data Acquisition

For our experiment, we have used the *myPersonality* data set [4], [31] that consists of the status updates, SN features, ground-truth personality traits scores, as well as classes. The traits used in the data set are formalized in BFFM. For each of the five personality traits, openness to experience (OPN), conscientiousness (CON), extraversion (EXT), agreeableness (AGR), and neuroticism (NEU), the personality score and the class value (yes or no) are given in the data set.

The data set contains 250 users around 10 000 status updates, and it is considered as a ground-truth data set for personality prediction. The class distribution of the *myPersonality* data set is demonstrated in Table I.

B. Data Preprocessing

All the statuses of the data set are in English and follow every step of preprocessing. The preprocessing step consists of the removal of URLs, names, symbols, unnecessary spaces, and stemming. These operations are performed using the NLTK package [45] library.

TABLE II
TRADITIONAL LINGUISTIC FEATURES

Feature No.	Feature Description
Character-level Features	
F1	No. of Characters
F2	No. of Punctuations
F3	No. of Special Characters
F4…F29	No. of individual alphabets (a, b, c,…z)
F30	Total no. of Alphabets
Word-level Features	
F31	No. of Words
F32	No. of words with 1 character
F33	No. of words with 2 character
F34	No. of words with 3 character
F35	No. of words with 4 character
F36	No. of words with 5 character
F37	No. of words with 6 character
F38	No. of words with 7 character
F39	No. of words with 8 character
F40	No. of words with 9 character
F41	No. of words with 10 character
F42	No. of words with 11 character
F43	No. of words with 12 character
F44	No. of words more than 12 character
F45	Avg. Word Length
Structural Features	
F46	No. of Sentence
F47	Avg. Sentence Length in terms of Character
F48	Avg. Sentence Length in terms of words
Function Words	
F49	No. of Function Words
F50	Percentage of Noun
F51	Percentage of Pronoun
F52	Percentage of Verb
F53	Percentage of Adjective
F54	Percentage of Adverb
F55	Percentage of Preposition
F56	Percentage of Conjunction
F57	Percentage of Interjection

C. Feature Extraction

In this step, the extracted features are in two categories: linguistic features and SN features. We have extracted the traditional linguistic features and psycholinguistic features as well.

1) Traditional Linguistic Features: The traditional linguistic features are textual features that could be divided into four types: character-based, word-based, structural, and function words. The list of traditional features considered for our study is shown in Table II. For extracting the linguistic features, we have applied LIWC [16] on the preprocessed textual data. LIWC gives a total of 93 features having psycholinguistic and traditional linguistic categorical features. All the features are integer or fractional values, meaning the percentages of words in specific categories.

2) Psycholinguistic Features: Among 93 features, only 28 could be considered as psycholinguistic features divided into five categories, namely, emotional affect, cognitive process, self-focus, social relationships, and perceptions.

TABLE III
PSYCHOLINGUISTIC FEATURES EXTRACTED USING LIWC

Feature No.	Feature Description
Emotional Affect	
F58	Affect
F59	Positive emotion
F60	Negative emotion
F61	Anxiety
F62	Anger
F63	Sad
Cognitive Process	
F64	Cognitive process
F65	Insight
F66	Cause
F67	Discrepancy
F68	Tentative
F69	Certain
F70	Different
Social Relationships	
F71	Social
F72	Family
F73	Friend
F74	Female
F75	Male
Self-focus	
F76	Self-focus
F77	Work
F78	Leisure
F79	Home
F80	Money
F81	Religion
F82	Death
Perceptions	
F83	perception
F84	See
F85	Feel
F86	Hear

The features associated in each of the categories are demonstrated in Table III.

Apart from the psycholinguistic features, another 65 different linguistic features are extracted using LIWC. The linguistic features are word count, analytical word, tone, word per sentence, number of six-letter words, number of articles, different punctuation symbols (period, comma, colon, semicolon, question mark, exclamatory mark, dash, quote, apostrophe, parenthesis, and so on), and so on. The percentage of function words or parts of speech, such as percentage of a noun, pronoun (personal and impersonal pronoun), preposition, adverb, conjunction, verb, adjective, comparative, and interrogative-words are considered as extracted linguisitc features.

3) Social Network Features: The second type of feature category is SN features. In SNSs, the architecture is build upon a graph. Each of the users is considered as one of the nodes of this huge graph. The edge between these nodes could be considered as the friend or connection between users. Therefore, SN works as a huge graph. Similar gene network analysis could be utilized through ranking of biomolecules for biomedical data sets [56]. Moreover, in the *myPersonality* data set, SN features are extracted from this huge graph. The SN features are network size (*F87*), betweenness (*F88*), n-betweenness (*F89*), density (*F90*), brokerage (*F91*), n-brokerage (*F92*), and transitivity (*F93*). These features are closely related to the behavior and personality of a user.

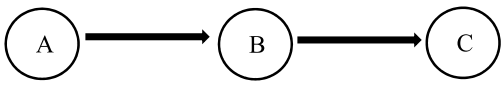

Fig. 2. Broker B between A and C.

Network size defines the number of friends, connections, or followers in SNSs. Using this feature, we may predict if the user has a decent number of friends or not. Having a smaller number of friends may lead to the characteristics of introvert user and vice versa

$$\text{NS}(v) = \text{Total no. of edges of } v. \tag{8}$$

Betweenness centrality $g(v)$ of a node v in a given graph could be determined using (7). Centrality is the measure to determine the central nodes within a graph, whereas the betweenness centrality demonstrates how many times a node behaved as a connector along the shortest path between two other nodes [54]. This measure is useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate neighborhood

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}. \tag{9}$$

Normalized betweenness centrality is the normalized value of $g(v)$ with respect to the minimum and maximum values of g [54], [55]. The formula to determine the n-betweenness is (3)

$$\text{normal}(g(v)) = \frac{g(v) - \min(g)}{\max(g) - \min(g)}. \tag{10}$$

Density is the measure of network connections. Network density could be measured using the formula (4). Density demonstrates the potential connections in a network, which are actual connections

$$\text{Density} = \frac{\text{Actual Connections}}{\text{Maximum Possible Connections}}. \tag{11}$$

Brokerage refers to the nodes embedded in its neighborhood, which is very useful in understanding power, influence, and dependence effects on graphs. A broker could be considered as the communicator between two different nodes [55]. Five types of brokers are available in the literature, namely, coordinator, consultant, gatekeeper, representative, and liaison. It is possible that different types of brokers are present in a simple SN graph. The general concept of brokerage could be depicted as in Fig. 2.

In a graph, if A is connected to B, and B is connected to C, but A and C are not connected to each other, and then A needs B to communicate with C. Thus, B is the broker node here.

The description of five different types of broker nodes is illustrated in Table IV. The equations used in Table IV are considering node B as a broker, and G(X) denotes the group that node x belongs to. It is presumed that A→ B→ C, thus

TABLE IV
DIFFERENT CATEGORIES OF BROKER NODE

Type of Broker	Equation to calculate brokerage	Conditions/ Criteria
Coordinator	Counts the no. of times B is a broker and G(A) = G(B) = G(C)	All three nodes belong to the same group
Consultant	Counts the no. of times B is a broker and G(A) = G(C), but G(B) ≠ G(A)	The broker belongs to one group and the other belong to a different group
Gatekeeper	Counts the no. of times B is a broker and G(A) ≠ G(B) and G(B) = G(C)	The source node belongs to a different group
Representative	Counts the no. of times B is a broker and G(A) = G(B) and G(C) ≠ G(B)	The destination node belongs to a different group
Liaison	Counts the no. of times B is a broker and G(A) ≠ G(B) ≠ G(C)	Each node belongs to a different group

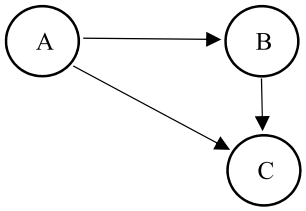

Fig. 3. Idea of transitivity.

A be the source node gives information to B, the broker node, who gives the information to C (the destination node).

N-brokerage is the normalized parameter of brokerage that is the measure of brokerage nodes divided by the number of pairs [55]. The equation could be derived as follows:

$$\text{n-brokerage} = \frac{\text{number of broker nodes}}{\text{number of pairs}}. \tag{12}$$

Transitivity is the measurement that could be defined as a friend-of-friend (FOF) concept of social media, such as Facebook. The idea of FOF is "when a friend of my friend is my friend."

In the context of network or graph theory, transitivity is measured based on the relative number of triangles or triads present in the graph comparing to the total number of connected triples of nodes. The idea of transitivity is depicted in Fig. 3, and the equation to calculate the transitivity $T(G)$ is (13), as follows.

As shown in Fig. 3, A is friend with B, B is friend with C, and A is also friend with C. The relationships between them build a triad

$$T(G) = \frac{3 * \text{no. of closed triples in } G}{\text{no. of connected triples of vertices in } G}. \tag{13}$$

TABLE V
PERFORMANCE MATRICES APPLYING CLASSIFIERS FOR PREDICTING EXTRAVERSION USING LIWC FEATURES

	Precision	Recall	F-Measure	ACC	MCC	ROC Area	PRC Area
NB	0.541	0.455	0.426	45.49%	0.027	0.513	0.548
DT	0.513	0.525	0.518	52.46%	-0.023	0.47	0.506
RF	0.57	0.607	0.541	60.66%	0.074	0.539	0.561
SLR	0.457	0.574	0.47	57.38%	-0.086	0.465	0.519
SVM	0.373	0.611	0.463	**61.07%**	0.024	0.5	0.524
SD	0.078	0.065	0.046	0.065	0.06	0.03	0.022

Bold values indicate the highest accuracy found for extraversion traits

D. Feature Selection

Feature selection algorithms are used to find the essential or important features from a set of the feature vector. The experiment starting from inputting the labeled data, data pre-processing, is performed and left for identifying the features. In features extraction step, we have collected the prominent features, each feature vector containing 93 features

$$F = \{F1, F2, F3, \ldots, F93\}. \tag{14}$$

These feature vectors are used to find the optimal number of essential features using the features selection methods. The mentioned five different types of feature selection methods are adopted, and the selected features are feed to the classifiers. The performance matrices are determined for each classifier to evaluate the experimental method. Finally, the highly accurate feature selection algorithm is identified.

E. Classification Methods

In this article, we have applied the classic classification methods to evaluate the performance of the proposed experimental method. NB, RF, DT, SLR, and SVM are implemented in the experiment process. The state-of-the-art classifiers are considered for the experiment, not adaptive versions.

VI. EXPERIMENTAL RESULTS

In this section, we have evaluated the system using the evaluation metrics, such as precision, recall, f-measure, and accuracy. We have divided the research contributions into three different experiments.

1) Experiment 1 (Using All LIWC Features for Predicting Five Personality Traits): In this experiment, we have focused on the 93 LIWC features to fetch into the prediction model. Table V shows the metrics values along with the SD applying different classifiers for LIWC features for extraversion trait only. The highest accuracy (61.07%) is shown by the SVM classifier for the extraversion trait. The same metrics are reported for the other four personality traits in Table VI.

TABLE VI
ACCURACY MEASUREMENTS OF EXPERIMENT-1 ON OCEAN TRAITS

Traits	NB	DT	RF	SLR	SVM	AVG	SD
O	46.72%	59.84 %	**69.26%**	68.85 %	67.62 %	62.46%	0.0959
C	52.62%	51.23 %	47.54%	53.69 %	**54.51 %**	51.92%	0.0273
E	45.49%	52.46 %	**60.66%**	57.38 %	61.07 %	55.41%	0.0653
A	**57.79%**	47.13 %	51.23%	51.64 %	54.92 %	52.54%	0.0403
N	**60.66%**	59.02 %	58.20%	59.84 %	59.86 %	59.52%	0.0093
AVG	52.66%	53.94 %	57.38%	58.28 %	59.60 %		
SD	0.0665	0.0539	0.0847	0.0671	0.0534		

Bold values indicate the highest accuracy found for that specific personality traits

For each of the personality traits, the accuracy along with the average and SD is tabulated in Table VI. The average and SD of the classifiers for each trait are given in the last two columns, and the average and SD of each classifier are given in the last two rows. The highest accuracy reported for each trait is kept in bold.

2) Experiment 2 (Using Psycholinguistic Cues and Feature Selection Algorithms Applied for Predicting Five Personality Traits): For this experiment, we have focused on the psycholinguistic features mostly and the combination of SN features. The performance metrics are determined for all the combinations and personality traits. We have compared the cases with and without SN features. The SN features are proved to be closely related to the class, as in each case, the accuracy measurement is higher than without using these features. Table VII demonstrates the accuracy along with the SD in all the scenarios for extraversion trait only. In this article, for experimental analysis, we have considered extracting more than 100 feature sets and the used combination of these features to find the best set of features using features selection algorithms.

It is evident in the literature review that various types of features generate quite different results for the prediction system. Therefore, we have tried 15 different combinations with and without applying features selection algorithms. In Table VII, we have listed the scenarios that we have considered. The first ten feature scenarios are without applying the feature selection algorithms, and the last five feature combinations are extracted applying five different feature selection algorithms. It is found in experiments that according to diversified psycholinguistic features, traditional classifiers are acting differently.

These features' combination scenarios are considered separately for each of the five personality traits, and the performance matrices are determined. As stated in different state-of-the-art articles and for ease of understanding, we are comparing the scenarios based on the accuracy (ACC).

As stated in Table VII, the highest accuracy is obtained from the Pearson correlation-based feature selection algorithm. We have determined the feature-class correlation index for each of the features and only selected the features having $\rho > 0.10$, therefore the higher correlated features. The 17 selected features using this method are all seven SN features (network size, betweenness, n-betweenness, density, brokerage, n-brokerage, and transitivity) and ten LIWC features (pronoun, they, I, filler, drives, authentic, dash, interrog, reward, and body).

TABLE VII
ACCURACY MEASUREMENTS APPLYING CLASSIFIERS FOR PREDICTING EXTRAVERSION IN DIFFERENT SCENARIOS

	NB	DT	RF	SLR	SVM
All Psycholinguistic Cues	47.54%	61.07%	57.79%	61.07%	61.07%
LIWC + Social Network (SN) Features	53.69%	56.97%	68.03%	63.52%	61.07%
Only SN Features	58.61%	65.98%	68.85%	67.62%	67.62%
Emotion Affect + SN Features	57.38%	65%	66%	66%	68%
Cognitive Process + SN Features	57.38%	63.52%	65.16%	65.98%	68.03%
Function Words + SN Features	58.61%	61.48%	67.62%	68.85%	69.26%
Self-Focus + SN Features	56.56%	64.34%	66.80%	66.80%	67.21%
Social Relationships + SN Features	69.67%	63.52%	65.16%	65.57%	68.03%
Perceptions + SN Features	69.26%	65.98%	63.93%	67.21%	67.62%
CFS Subset based Selected Features	59.43%	66.80%	67.21%	66.39%	65.16%
Information Gain (IG) based Selected Features	58.61%	65.95%	68.85%	67.62%	67.62%
Symmetric Uncertainty (SU) Based Selected Features	57.38%	65.16%	68.03%	66.39%	67.62%
Chi-Squared Based Selected Features	61.07%	65.17%	64.75%	67.62%	67.21%
Pearson Correlation based Selected Features	**72.13%**	62.70%	59.26%	64.34%	69.67%
SD	7.328	3.874	3.468	2.954	2.937

Bold value indicates the highest accuracy found for all the feature combinations.

The comparative scenario of using the SN and not using the SN features with the individual psycholinguistic cues are given in Fig. 4. Fig. 4 illustrates that the SN features have deep insight and influential factors, as for each of the cases (except for NB and cognitive process), input features having SN features are giving better accuracy. Therefore, for each of

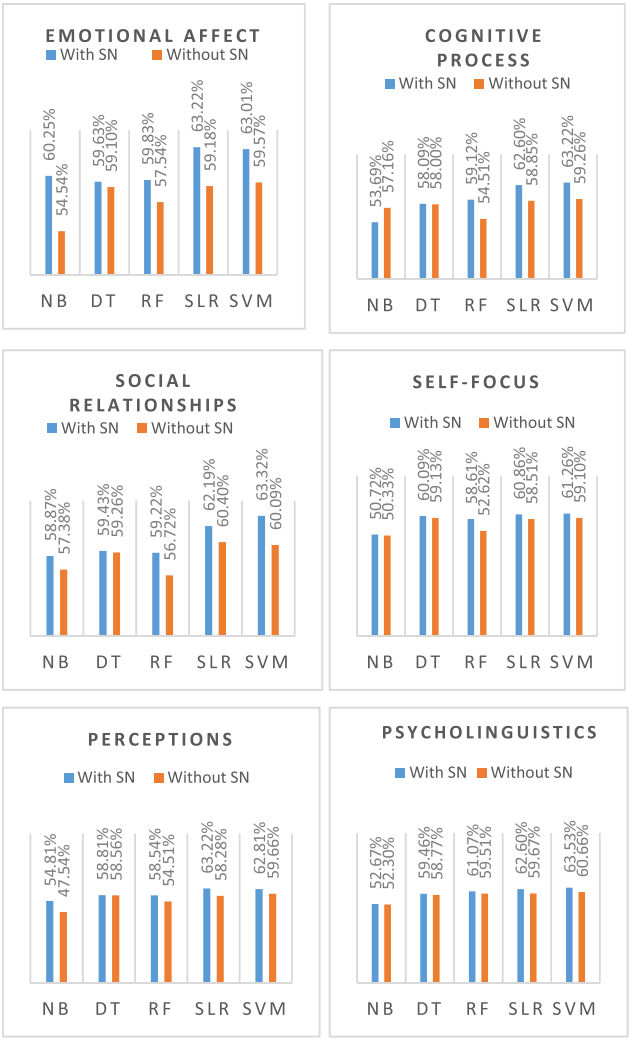

Fig. 4. With and without SN features' accuracy measures of classifiers.

the personality traits, it is evident that SN features are selected while using feature selection algorithms.

VII. Comparative Analysis

A. Comparisons on Feature Combination Scenarios

As demonstrated in Table VII, for all 15 feature combination scenarios, five classifiers have been applied to predict the big-five-personality traits. We have compiled the best/highest accuracy (kept in bold) for predicting each of the traits and illustrated in Table VIII. In Table VIII, the number in the brackets gives the classification method number, where NB is (1), DT is (2), and RF, SLR, and SVM are (3), (4), and (5), respectively. For example, LIWC features give 69.67% accuracy using NB for predicting openness-to-experience. For this article, we have concentrated on the last five features' combination scenarios focusing on the feature selection algorithms. The data in Table VIII visualize that the Pearson correlation-based selected features are always giving the highest accuracy for each of the personality traits. The classification algorithm used for getting the highest accuracy is NB, except for the openness-to-experience trait (O). RF algorithm gives the highest accuracy for the openness trait.

TABLE VIII

Accuracy Measurements Applying Classifiers for Predicting Big-Five-Personality Traits

	Openness-to-Experience (O)	Conscientiousness (C)	Extraversion (E)	Agreeableness (A)	Neuroticism (N)
LIWC Features	69.67% (1)	54.51% (5)	61.07% (5)	57.79% (1)	59.84% (3)
Psycholinguistic Cues	69.66% (5)	58.61% (1)	61.07% (2, 4, 5)	54.51% (5)	61.89% (1)
LIWC + Social Network (SN) Features	68.25% (3)	54.92% (5)	68.03% (3)	57.79% (4)	58.20% (1)
Only SN Features	69.67% (1)	58.20% (4)	68.85% (3)	57.38% (5)	62.70% (4)
Emotion Affect + SN Features	65.57% (3)	61.07% (4)	68% (5)	59.84% (5)	66.39% (3)
Cognitive Process + SN Features	67.21% (3)	58.61% (4)	68.03% (5)	58.20% (5)	62.30% (3)
Function Words + SN Features	69.23% (1)	55.33% (4)	69.26% (5)	56.56% (5)	63.52% (3)
Self-Focus + + SN Features	65.87% (3)	52.46% (4)	67.21% (5)	57.38% (5)	63.11% (1)
Social Relationships + SN Features	65.17% (3)	57.79% (4)	69.67% (1)	59.02% (5)	63.11% (3)
Perceptions + SN Features	68.98% (3)	59.02% (4)	69.26% (1)	57.79% (5)	64.16% (4)
CFS Subset based Selected Features	69.67% (3)	59.02% (3)	66.80% (3)	56.56% (1)	63.16% (1)
Information Gain (IG) based Selected Features	68.27% (5)	59.02% (3)	68.85% (3)	56.56% (1)	63.52% (4)
Symmetric Uncertainty (SU) Based Selected Features	67.67% (5)	59.02% (3)	68.03% (3)	60.25% (5)	61.48% (4)
Chi-Squared Based Selected Features	69.67% (5)	60.65% (3)	67.63% (3)	60.66% (5)	63.52% (4)
Pearson Correlation based Selected Features	**70.08% (3)**	**66.39% (1)**	**72.13% (1)**	**61.89% (1)**	**64.75% (1)**
SD	1.660	3.272	2.930	1.903	1.948

Bold value indicates the highest accuracy found for each of the personality traits

For comparing the feature selection methods in Table IX, the pairwise t-test is reported for showing the statistical significance of the methods, as shown in [47]. For simplicity, we have reported the t-test values and degree of freedom values of the PCC method with the rest of the methods for

TABLE IX
T-TEST RESULTS FOR FEATURE SELECTION METHODS

Properties	CFS	IG	SU	CHI
t-test	1.32218	1.24530	1.42170	1.01158
df	7	7	8	8

TABLE X
WIN-DRAW-LOSS TABLE OF FEATURE SELECTION METHODS

FS Methods	CFS	IG	SU	CHI	PCC
CFS	--	1-2-2	2-1-2	0-1-4	0-0-5
IG	2-2-1	--	3-1-1	1-1-3	0-0-5
SU	2-1-2	1-1-3	--	1-0-4	0-0-5
CHI	4-1-0	3-1-1	4-0-1	--	0-0-5
PCC	5-0-0	5-0-0	5-0-0	5-0-0	--

accuracy metric. The two-tailed t-test is performed assuming alpha = 0.05, and the hypothesis means difference is equal to zero.

From Table VIII, we can claim that among the feature selection methods, PCC is providing better accuracy than the others. For a better understanding of the comparison, we have computed the win–draw–loss as constructed in [48]–[50] into Table X describing how many times each method has won against the other methods. Form the data, we can see that PCC has won against all other methods each time and not even drawn with any method. Therefore, we had analyzed the features selected using the PCC method, in Section VII-B.

B. Insights of the Pearson Correlation-Based Selected Features

From the PCC calculation, as given in (7), is invariant under separate modifications in scale and location in the two variables, which could be considered as a key mathematical property of PCC. Therefore, the PCC has been used in diversified research problem for the same purpose of feature selection. Kim *et al.* [62] presented a correlation analysis for DNA microarray data sets, such as leukemia, colon, and lymphoma. They utilized the ensemble classifiers to get the highest accuracy on each of the data sets. PCC has been utilized in image processing, such as tissue classification from CT images [63].

The implication of PCC for noise removal in the context of signal processing is presented in [64]. They provide experimental justification for using PCC on signal data. The statistical perspective of using PCC has been presented in [64], which focuses on the medical research domain. A practical application of PCC has been demonstrated in [64] using the sample data of 780 women attending their first antenatal clinic visits. In the context of natural language processing (NLP), PCC has proven to work better for many applications, such as neurolinguistic approach to NLP using medical text analysis [65], automated classification of radiology reports for acute lung injury using machine learning and NLP [66], finding strong correlation between text quality and complex network features [67], detection using NLP [68].

The selected features that are determined by applying the Pearson correlation-based feature selection method give very promising insights about personality traits. Here, we have considered the features in set representation and found interesting combinations for different traits. For each of the traits, the sets are named using their initial, such as E for extraversion and N for Neuroticism.

E = {network-size, betweenness, n-betweenness, density, brokerage, n-brokerage, transitivity, ppronoun, they, I, filler, drives, Authentic, Dash, interrog, reward, body}

N = {network-size, betweenness, density, brokerage, transitivity, relig, number, comma, differ, work}

A = {parenth, transitivity, clout, we, social, nonflu, they, adverb, n-betweenness, swear, quote, informal, she/he, word-per-sentence (WPS), differ, male}

C = {network-size, betweenness, n-betweenness, density, brokerage, sad, Dash, friend, social, feel, clout, you, colon, power, authentic, Dic, percept, male, family, anx, affiliation, differ, discrep}

O = {informal, feel, affect, conj, filler, focuspast, swear, allpunc, period}

$U = E \cup N \cup A \cup C \cup O$

= {network-size, betweenness, n-betweenness, density, brokerage, n-brokerage, transitivity, ppronoun, they, I, filler, drives, Authentic, Dash, interrog, reward, body, relig, number, comma, differ, work, parenth, clout, we, social, nonflu, adverb, swear, quote, informal, she/he, word-per-sentence(WPS), male, sad, friend, feel, you, colon, power, Dic, percept, family, anx, affiliation, discrep, conj, focuspast, swear, allpunc, period}

$E \cap N$ = {network-size, betweenness, density, brokerage, transitivity}

$E \cap A$ = {n-brokerage, transitivity}

$E \cap C$ = {network-size, betweenness, n-betweenness, density, brokerage, Dash}

$E \cap O = \{\emptyset\}$.

From the above-mentioned sets, we can depict that the SN features are playing an influential role in high-accuracy predictions. The seven SN features could be found in each of the traits sets showing the influence except for set O. Therefore, openness-to-experience (O) trait has lesser correlations with the SN features. The universal (U) set represents the set having union of all the sets, which contains 51 distinct features. From the selected features from the Pearson correlation, we have got the highest accuracy of 72.13% applying NB classifier for extraversion trait. From the above-mentioned sets, we can declare the following findings.

1) SN features are the most prominent features as they are highly correlated with personality traits.
2) Among the psycholinguistic features, all the "social relationship" features are found in the universal set except the number of female-related words.

TABLE XI
COMPARISON WITH THE LITERATURE METHODS

Features Used	Feature Selection Used	Classification Method Used	Evaluation
Linear Dirichlet Allocation (LDA) and Term Frequency-Inverse Document Frequency (TF-IDF) [12]	No	Support Vector Regression (SVR) SVR-Linear, Poly, RBF Decision Tree	MSE= 0.0017 (SVR) for Conscientiousness
Activity & demographic information, SentiStrength [46]	No	Linear Regression, SVM	RMSE-0.651 Using all features for Openness
Time-based and social network feature [61]	Yes (Manual fusion)	Support Vector Machine (SVM), kNN and NB	Accuracy: 63% for SVM and kNN Extraversion Trait
LIWC, SPLICE and SN features [13]	Yes (Manual fusion)	Deep learning algorithms MLP, CNN-1D, LSTM, GRU	Accuracy: 70.78% (MLP) for myPersonality Accuracy: 74.17% (LSTM + CNN-1D) for Bahasa
Proposed Method (Traditional linguistic, psycholinguistic and social network features)	Yes (Manual and Automated feature selection) FS Algorithms: CFS, IG, SU, CHI, PCC	NB, DT, RF, SLR & SVM	Accuracy: 72.13% Extraversion trait

3) The influence of punctuations are at a decent level, such as dash, comma, parenth, quotes, colon, period, and allpunc features are present in the universal set.
4) The important function words (personal pronoun, interrogative, adverb, and conjunction) are present in the U set and have a good correlation with the relative classes.
5) The openness-to-experience trait has shown divert results, and the selected features' set does not contain any SN features.
6) PCC outperforms the other existing feature selection algorithms for predicting personality from social media using linguistic and SN features.

C. Comparison With Literature Methods

However, there are very few works found in the literature utilizing the traditional linguistic, psycholinguistic, and SN features altogether for predicting personality from social media. Table XI shows comparisons with the literature methods and features used for predicting with our approach. The proposed feature selection approach in this article has shown better accuracy using the selected features through the PCC algorithm.

VIII. CONCLUSION

We have presented a comparative analysis among the feature or attribute selection algorithms for predicting personality using positive and negative traits. This article works with the user-generated social media contents, such as Facebook status updates, and extracted the most relevant features, including the textual features, such as traditional and psycholinguistic features.

As we know, the base of social media is basically a graph. The connections between the nodes and impact on them due to social media interactions could be reflected through the SN features. We have designed and performed experiments utilizing the linguistic features as well as SN features. To the best of our knowledge, we have used the most number of features to compare the performance of the feature selection algorithms to predict personality traits. Feature combinations or subset-based scenarios are used to identify the best possible features. According to the experimental findings, among the tested algorithms, PCC-based selected features have outperformed the literature methods giving 72.13% accuracy for extraversion trait. The overall accuracy of each of the personality traits of BFFM has increased after using PCC-based feature selection algorithm.

REFERENCES

[1] M. M. Hasan, N. H. Shaon, A. A. Marouf, M. K. Hasan, H. Mahmud, and M. M. Khan, "Friend recommendation framework for social networking sites using user's online behavior," in *Proc. 18th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2015, pp. 539–543.
[2] M. S. H. Mukta, M. E. Ali, and J. Mahmud, "User generated vs. Supported contents: Which one can better predict basic human values?" in *Proc. Int. Conf. Social Inform.* Cham, Switzerland: Springer, 2016, pp. 454–470.
[3] C. P. Williams. (Feb. 23, 2013). *Language, Identity, Culture, and Diversity*. [Online]. Available: https://www.newamerica.org/education-policy/edcentral/ multilingualismmatters/
[4] 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. Psycholog.*, vol. 70, no. 6, pp. 543–556, Sep. 2015.
[5] *International Personality Item Pool*. Accessed: Jan. 7, 2020. [Online]. Available: https://ipip.ori.org/
[6] 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. (WebSci)*, Evanston, IL, USA, Jun. 2012, pp. 24–32.
[7] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *Proc. Extended Abstracts Hum. Factors Computing Syst.*, Vancouver, BC, Canada, May 2011, pp. 253–262.
[8] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski, and J. Crowsroft, "The personality of popular Faebook users," in *Proc. CSCW*, Seattle, WA, USA, Feb. 2012, pp. 955–964.
[9] K. Moore and J. C. Mcelroy, "The influence of personality on Facebook usage, wall postings, and regret," *Comput. Human Behavior*, vol. 28, no. 1, pp. 267–274, Jan. 2012.
[10] A. Ortigosa, R. M. Carro, and J. I. Quiroga, "Predicting user personality by mining social interactions in Facebook," *J. Comput. Syst. Sci.*, vol. 80, no. 1, pp. 57–71, Feb. 2014.
[11] A. Eftekhar, C. Fullwood, and N. Morris, "Capturing personality from Facebook photos and photo-related activities: How much exposure do you need?" *Comput. Hum. Behav.*, vol. 37, pp. 162–170, Aug. 2014.
[12] P. Howlader, K. K. Pal, A. Cuzzocrea, and S. D. M. Kumar, "Predicting Facebook-users' personality based on status and linguistic features via flexible regression analysis techniques," in *Proc. 33rd Annu. ACM Symp. Appl. Comput. (SAC)*, Pau, France, Apr. 2018, pp. 339–345.
[13] T. Tandera, D. Suhartono, R. Wongso and Y. L. Prasetio, "Personality preddiction system from Facebook users," in *Proc. 2nd Int. Conf. Comput. Sci. Comput. Intell. (ICCSCI)*, Bali, Indonesia, Oct. 2017, pp. 604–611.

[14] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell, "Mining Facebook data for predictive personality modeling," Comput. Pers. Recognit., AAAI, Tech. Rep., Menlo Park, CA, USA, 2013, pp. 23–26.

[15] V. Kaushal and M. Patwardhan, "Emerging trends in personality identification using online social networks—A literature survey," *ACM Trans. Knowl. Discov. DataT*, vol. 12, no. 2, pp. 1–30, Jan. 2018.

[16] J. W. Pennebaker, M. E. Francis, and R. J. Booth. (2001). Linguistic Inquiry and Word Count: LIWC2001. Erlbaum, Mahwah, NJ, USA. [Online]. Available: https://www.erlbaum.com

[17] M. Coltheart, "The MRC psycholinguistic database," *Quart. J. Exp. Psychol. A*, vol. 33, no. 4, pp. 497–505, Nov. 1981.

[18] K. Moffitt, J. Giboney, E. Ehrhardt, J. Burgoon, and J. Nunamaker. (2010). *Structured Programming for Linguistic CUE Extraction*. [Online]. Available: http://splice.cmi.arizona.edu/

[19] Kucera and W. N. Francis, *Computational Analysis of Present-day American English*. Providence, RI, USA: Brown Univ. Press, Providence, 1967.

[20] G. D. A. Brown, "A frequency count of 190,000 words in the London-Lund Corpus of English conversation," *Behav. Res. Methods Instrum. Comput.*, vol. 16, no. 6, pp. 502–532, 1984.

[21] L. R. Goldberg, "The development of markers for the big-five factor structure," *Psychol. Assessment*, vol. 4, no. 1, pp. 26–42, 1992, doi: 10.1037/1040-3590.4.1.26.

[22] I. B. Myers, M. H. McCaulley, N. L. Quenk, and A. L. Hammer, *MBTI Manual: A Guide to the Development and Use of the Myers–Briggs Type Indicator*, vol. 3, 3rd ed. Palo Alto, CA, USA: Consulting Psychologists Press, 1998.

[23] W. Marston, *Emotions of Normal People*. New York, NY, USA: Taylor & Francis, 1999.

[24] P. T. Costa, Jr., and R. R. McCrae, *NEO-PI-R Professional Manual*. Odessa, FL, USA: Psychological Assessment Resources, 1992.

[25] L. R. Goldberg, "A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models," in *Personality Psychology in Europe*, vol. 7, I. Mervielde, I. Deary, F. De Fruyt, and F. Ostendorf, Eds. Tilburg, The Netherlands: Tilburg Univ. Press, 1999, pp. 7–28. [Online]. Available: http://ipip.ori.org/newBroadbandText.htm

[26] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of Personality: Theory and Research*, L. A. Pervin and O. P. John, Eds., 2nd ed. New York, NY, USA: Guilford Press, 1999, pp. 102–138.

[27] G. Saucier, "Mini-markers: A brief version of Goldberg's unipolar big-five markers," *J. Pers. Assessment*, vol. 63, no. 3, pp. 506–516, Dec. 1994.

[28] M. B. Donnellan, F. L. Oswald, B. M. Baird, and R. E. Lucas, "The mini-IPIP scales: Tiny-yet-effective measures of the big five factors of personality," *Psychol. Assessment*, vol. 18, no. 2, pp. 192–203, Jun. 2006.

[29] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the Big-Five personality domains," *J. Res. Pers.*, vol. 37, no. 6, pp. 504–528, 2003.

[30] J. A. Johnson, "Developing a short form of the IPIP-NEO: A report to HGW Consulting," Dept. Psychol., Univ. Pennsylvania, DuBois, PA, USA, Tech. Rep., 2000.

[31] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 15, pp. 5802–5805, Apr. 2013.

[32] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 4, pp. 1036–1040, Jan. 2015.

[33] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, Apr. 1999.

[34] M. Hall and L. A. Smith, "Feature subset selection: A correlationbased filter approach," in *Proc. 4th Int. Conf. Neural Inf. Process. Intell. Inf. Syst.*, 1997, pp. 855–858.

[35] M. Doshi and R. K. Chaturvedi, "Correlation based feature selection (CFS) technique to predict student perfromance," *Int. J. Comput. Netw. Commun.*, vol. 6, no. 3, pp. 197–206, May 2014.

[36] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 447–462, Mar. 2011.

[37] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003.

[38] Z. Gao, Y. Xu, F. Meng, F. Qi, and Z. Lin, "Improved information gain-based feature selection for text categorization," in *Proc. 4th Int. Conf. Wireless Commun., Veh. Technol., Inf. Theory Aerosp. Electron. Syst. (VITAE)*, May 2014, pp. 11–14.

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

[40] E. Sarhrouni, A. Hammouch, and D. Aboutajdine, "Application of symmetric uncertainty and mutual information to dimensionality reduction of and classification hyperspectral images," *Int. J. Eng. Technolofy*, vol. 4, no. 5, pp. 268–276, 2012.

[41] S. Imranali and W. Shahzad, "A feature subset selection method based on symmetric uncertainty and ant colony optimization," *Int. J. Control Automat.*, vol. 60, no. 11, pp. 5–10, Jul. 2017.

[42] K. Pearson, *On the Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables is Such That It Can Be Reasonably Supposed to Have Arisen From Random Sampling*. London, U.K.: Philosophical Magazine, vol. 5, no. 50. 1900, pp. 157–175, doi: 10.1080/14786440009463897.

[43] M. S. Nikulin, "Chi-squared test for normality," in *Proc. Int. Vilnius Conf. Probab. Theory Math. Statist.*, vol. 2, 1973, pp. 119–122.

[44] B. Jacob, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction Speech Processing*. Berlin, Germany: Springer-Verlag, 2009, pp. 1–4.

[45] E. Loper and S. Bird, "NLTK: The natural language toolkit," in *Proc. ACL Workshop Effective Tools Methodologies Teach. Natural Lang. Process. Comput. Linguistics*, 2002, pp. 1–8.

[46] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, and M. De Cock, "Computational personality recognition in social media," *User Model. User-Adapted Interact.*, vol. 26, nos. 2–3, pp. 109–142, 2016.

[47] S. Bandyopadhyay, S. Mallik, and A. Mukhopadhyay, "A survey and comparative study of statistical tests for identifying differential expression from microarray data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 1, pp. 95–115, Jan. 2014, doi: 10.1109/tcbb.2013.147.

[48] S. Mallik and Z. Zhao, "ConGEMs: Condensed gene co-expression module discovery through rule-based clustering and its application to carcinogenesis," *Genes*, vol. 9, no. 1, p. 7, Dec. 2017, doi: 10.3390/genes9010007.

[49] S. Mallik, T. Bhadra, and U. Maulik, "Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based feature selection for multi-omics data," *IEEE Trans. Nanobiosci.*, vol. 16, no. 1, pp. 3–10, Jan. 2017, doi: 10.1109/tnb.2017.2650217.

[50] T. Bhadra, S. Mallik, and S. Bandyopadhyay, "Identification of multi-view gene modules using mutual information-based hypograph mining," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 6, pp. 1119–1130, Jun. 2019, doi: 10.1109/tsmc.2017.2726553.

[51] X. Xu, H. Gu, Y. Wang, J. Wang, and P. Qin, "Autoencoder based feature selection method for classification of anticancer drug response," *Frontiers Genet.*, vol. 10, p. 233, Jan. 2019, doi: 10.3389/fgene.2019.00233.

[52] S. Mallik and Z. Zhao, "Graph- and rule-based learning algorithms: A comprehensive review of their applications for cancer type classification and prognosis using genomic data," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bby120.

[53] S. Mallik and Z. Zhao, "Towards integrated oncogenic marker recognition through mutual information-based statistically significant feature extraction: An association rule mining based study on cancer expression and methylation profiles," *Quant. Biol.*, vol. 5, no. 4, pp. 302–327, Dec. 2017, doi: 10.1007/s40484-017-0119-0.

[54] (2019). *Mjdenny.com*. Accessed: Aug. 31, 2019. [Online]. Available: http://www.mjdenny.com/workshops/SN_Theory_I.pdf

[55] (2019). *unginstitute.berkeley.edu*. Accessed: Aug. 31, 2019. [Online]. Available: https://funginstitute.berkeley.edu/wp-content/uploads/2012/10/Brokerage-2C-Boundary-Spanning-2C-and-Leadership-in-Open-Innovation-Communities.pdf.

[56] S. Mallik and U. Maulik, "MiRNA-TF-gene network analysis through ranking of biomolecules for multi-informative uterine leiomyoma dataset," *J. Biomed. Informat.*, vol. 57, pp. 308–319, Oct. 2015, doi: 10.1016/j.jbi.2015.08.014.

[57] Y. Masoudi-Sobhanzadeh, H. Motieghader, and A. Masoudi-Nejad, "FeatureSelect: A software for feature selection based on machine learning approaches," *BMC Bioinf.*, vol. 20, no. 1, p. 170, 2019, doi: 10.1186/s12859-019-2754-0.

[58] M. Mafarja, I. Aljarah, H. Faris, A. I. Hammouri, A. M. Al-Zoubi, and S. Mirjalili, "Binary grasshopper optimisation algorithm approaches for feature selection problems," *Expert Syst. Appl.*, vol. 117, pp. 267–286, Mar. 2019, doi: 10.1016/j.eswa.2018.09.015.
[59] V. Chahkandi, M. Yaghoobi, and G. Veisi, "Feature selection with chaotic hybrid artificial bee colony algorithm based on fuzzy (CHABCF)," *J. Soft Comput. Appl.*, vol. 2013, pp. 1–8, Jun. 2013, doi: 10.5899/2013/jsca-00014.
[60] S. Arora and P. Anand, "Binary butterfly optimization approaches for feature selection," *Expert Syst. Appl.*, vol. 116, pp. 147–160, Feb. 2019, doi: 10.1016/j.eswa.2018.08.051.
[61] G. Farnadi, S. Zoghbi, M. Moens, and M. De Cock, "Recognising personality traits using Facebook status updates," in *Proc. WCPR*, 2013, pp. 14–18.
[62] K.-J. Kim and S.-B. Cho, "Ensemble classifiers based on correlation analysis for DNA microarray classification," *Neurocomputing*, vol. 70, nos. 1–3, pp. 187–199, Dec. 2006.
[63] B. Auffarth, M. Lopez-Sanchez, and J. Cerquides, "Comparison of redundancy and relevance measures for feature selection in tissue classification of CT images," *Advances in Data Mining. Applications and Theoretical Aspects*, P. Perner, Ed. Berlin, Germany: Springer, 2010, pp. 248–262.
[64] M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, Sep. 2012.
[65] W. Duch, P. Matykiewicz, and J. Pestian, "Neurolinguistic approach to natural language processing with applications to medical text analysis," *Neural Netw.*, vol. 21, no. 10, pp. 1500–1510, Dec. 2008.
[66] I. Solti, C. R. Cooke, F. Xia, and M. M. Wurfel, "Automated classification of radiology reports for acute lung injury: Comparison of keyword and machine learning based natural language processing approaches," in *Proc. IEEE Int. Conf. Bioinformatics Biomed. Workshop*, Washington, DC, USA, Nov. 2009, pp. 1–4.
[67] L. Antiqueira, M. Nunes, O. Oliveira, Jr., and L. D. F. Costa, "Strong correlations between text quality and complex networks features," *Phys. A, Stat. Mech. Appl.*, vol. 373, pp. 811–820, Jan. 2007.
[68] M. Chong, L. Specia, and R. Mitkov, "Using natural language processing for automatic detection of plagiarism," in *Proc. 4th Int. Plagiarism Conf.* Tyne, U.K.: Northumbria Univ. 2010, pp. 1–12.

Ahmed Al Marouf received the bachelor's degree from the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Gazipur, Bangladesh, in 2014, and the M.Sc.Engg. degree in CSE from IUT in 2019.

He was a Graduate Researcher with the Systems and Software Lab (SSL), CSE Department, IUT. He is currently a Lecturer with the Department of Computer Science and Engineering (CSE), Daffodil International University (DIU), Dhaka, Bangladesh, where he is also the Technical Lead of the Human Computer Interaction (HCI) Research Lab. His research interest lies within computational social science, data science, and machine learning.

Md. Kamrul Hasan received the B.Sc. degree in computer science and information technology (CIT) from the Islamic University of Technology (IUT), Gazipur, Bangladesh, and the Ph.D. degree from Kyung Hee University, Seoul, South Korea.

He has long experience in software as a developer and a consultant. He is currently a Professor with the Department of Computer Science and Engineering (CSE), IUT, where he has been serving for ten years and is also the Founding Director of the Systems and Software Lab (SSL). His current research interests are in intelligent systems and AI, software engineering, cloud computing, data mining applications, and social networking.

Hasan Mahmud received the bachelor's degree in computer science and information technology (CIT) from the Islamic University of Technology (IUT), Gazipur, Bangladesh, in 2004, and the M.Sc. degree in computer science from the University of Trento (UniTN), Trento, Italy in 2009. He is currently pursuing the Ph.D. degree in computer science and engineering (CSE) with IUT, under the guidance of Dr. M. A. Mottalib and Dr. K. Hasan.

He joined the CSE Department, Stamford University Bangladesh, Dhaka, Bangladesh, as a Faculty Member. Since 2009, he has been an Assistant Professor with the Department of CSE, IUT, where he is also the Co-Founder of the Systems and Software Lab (SSL). He has different research articles published in several international journals and conferences. His research interest focuses on human–computer interaction, gesture-based interaction, and machine learning.

Mr. Mahmud received the University Guild Grant Scholarship for two years (2007–2009) for his master's study and the Early Degree Scholarship.