International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 14 Issue 2 –APRIL 2015.

Social-Network-Sourced Big Data Analytics for Personality Prediction: A Review

Mayuri Pundlik Kalghatgi[#1], Manjula Ramannavar[*2], Dr. Nandini S.Sidnal [$3]

[#]*M.Tech, Dept. of CSE, KLS Gogte Institute of Technology, Belagavi, Karnataka, India*

[*]*Assistant Professor, Dept. of CSE, KLS Gogte Institute of Technology, Belagavi, Karnataka, India*

[$] *Professor, Dept. of CSE, KLE Dr. M. S. Sheshgiri College of Engg and Technology, Belagavi, Karnataka, India*

***Abstract*—The use of social networking websites has increased since last few decades. The social networking sites such as Twitter, Facebook, LinkedIn and YouTube allow users to create and share content related to different subjects, exposing their activities, feelings, thoughts and opinions. These sites have not only connected large user populations but have also captured massive information associated with their daily interactions in the form of Big Data. This data provides unprecedented information about human behavior and social interactions. It makes it possible to understand who the users are, what their interests are and what they need. This information is vital for a business to target potential consumers or seek customer opinions in the event of diversification as a business strategy. Thus, this paper reviews the techniques used for analyzing social media data to identify important personality traits, that is, characteristics or qualities particular to a person, which can be used in a variety of areas such as marketing, business intelligence, psychology and sociology. A parallelism among individual's personality traits and his/her linguistic information is explored for analytics.**

***Keywords*—Big Five model, Lexical Resources, Personality, Social Media.**

I. INTRODUCTION

Social networking on the web has become an essential component of everyday life. It has radically changed the ways in which people express their opinions and sentiments. Social networking sites such as Facebook, Twitter and YouTube are based on human interaction and the concept of user-generated content. This leads to the creation and exchange of a vast amount of user-generated content, entailing a massive production of free-form and interactive data [1]. Social media-oriented people tend to publish a lot about themselves through status updates, self-description, photos, videos and interests. The data available within social media platform is enormous in volume and reveals different aspects of human behavior and social interaction. An individual's personality is his/her characteristics and aspects that others can see. The data available on social media platform enables us to understand who the users are and what their needs are. Thus, the analysis of social media data makes it possible to identify important personality traits, that is, characteristics or qualities which describes his/her personality.

A parallelism among individual's personality traits and his/her linguistic information is obtained from the Big Five model. Computational model for predicting personality can be defined using social processes and family words. Such applications depend upon lexical techniques to predict personality from Twitter and Facebook data [2], [3]. Lexical analysis was used understand the meaning of words carrying sentimental or emotional content.

Developing a model that can accurately predict personality using social media texts has several applications. In marketing, it may be useful in identifying the sentiments hidden in a message e.g. for a product thereby revealing the likes or taste of an individual. This serves as a key factor for marketers who want to create an image in the minds of their customers for their product, brand, or organization and therefore identify which products to recommend to the user. In the field of psychology, it may be applied to social media data to understand user behavior; to study the dark triad (psychopathy, narcissism, and Machiavellianism) [3], [4], [5]; to identify criminal content; to model affection [6] etc.

Since an individual may normally have more than one personality trait, where each of these traits corresponds to a class for the classifier, the personality prediction based on the Big Five model can be considered as a multi-label classification. A multi-label classification is a classification problem where multiple target labels must be assigned to each instance. In the Big Five Model personality is divided into five dimensions namely Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism (OCEAN).

This work presents the techniques to predict personality based on the Big Five Model. The paper is organized as follows: Section II includes a discussion about the background of personality and Big Five model along with the data analysis and classification methods used for personality prediction; Section III deals with the work that has been carried out in this field; Section IV includes general discussion on strengths and limitations of reviewed approaches, and how they can be improved and section V concludes with suggestions for future research.

II. BACKGROUND

A. Personality and the Big Five Model

The term 'personality' is derived from the Latin word persona, which means the mask used by actors in a theatre. A set of attributes that characterize an individual and involves emotions, behavior, temperament and the mind defines a

TABLE I : BIG FIVE DIMENSIONS

Openness		Conscientiousness		Extroversion		Agreeableness		Neuroticism	
Low	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**
Commonplace	Wide interests	Careless	Organized	Quiet	Talkative	Fault-finding	Sympathetic	Stable	Tense
Simple	Imaginative	Disorderly	Tend to Plan	Reserved	Active	Cold	Kind	Calm	Anxious
Shallow	Intelligent	Frivolous	Efficient	Shy	Energetic	Unfriendly	Appreciative	Contented	Nervous
Unintelligent	Curious	Irresponsible	Responsible	Silent	Enthusiastic	Cruel	Generous	Unemotional	Worried

Source: Adapted from [18]

personality. Due to the diversity of attributes it is crucial to gauge personality as it does not provide any definitive structure through which people can be classified and compared. The set of human emotions is vast, due to which a similar problem occurs when one tries to identify the sentiment embedded in a message (sentiment analysis), thus making it challenging to choose the basic emotions for a classification. Thus in order to automate sentiment analysis, for instance, many researchers accepts a simplified representation of sentiments by means of their polarity (negative or positive) [7]. Similarly for determining personality, various researchers have recognized the most essential characteristics in order to create a personality model. Personality can vary depending on different situations. Thus, any personality prediction model must provide labels for all groups of characteristics. In analysis of the personality structure, definition of the Big Five Model or Five Factor Model came into use.

The "Big Five" model of personality dimensions is one of the well-experimented and well-scrutinized measures of personality structure used by researchers in recent years. The model describes a personality structure which is divided into five elements known as OCEAN: Openness, Conscientiousness, extroversion, Agreeableness, and Neuroticism, which were conceived by [8] as the key traits that emerged from investigation of previous personality tests [9]. Additional research proved that model's validity was not altered by different languages, tests and methods of analysis [9], [10]. Such broad research led many of psychologists to accept the Big Five model as the current definitive model of personality. Table I shows the five dimensions of the Big Five Model.

The Big Five traits can be described as follows:

- Openness: is ability of an individual to accept the new things. Individuals belonging to this category frequently use social media.
- Conscientiousness: indicates people who are meticulous, careful, punctual, thorough, and organized. Such people use less social media, because they believe that these sites serve as an unwanted distraction.
- Extroversion: indicates adventurous, sociable and talkative people. Such people tend to make lot of friends outside the virtual environments and invite them to the web, in order to keep in touch, but do not replace personal relationships.
- Agreeableness: indicates how friendly people are towards each other. Studies show that people with low levels of agreeableness might have a large number of online contacts but they find it difficult to initiate and maintain friendships beyond the virtual environment.
- Neuroticism: relate to control over the emotions. Such people use the Internet because they find it as a means of reducing loneliness and creating a sense of belonging.

B. Data analysis and Classification Methods

This section first expresses various lexical resources used to analyze the text messages and then explains classification methods used to train the classifiers.

1) Lexical Resources:

To identify personality associated with texts, information related to language and properties of individual words of concept is used. Particularly the following lexical resources are used.

a) LIWC (Linguistic Inquiry and Word Count):

Linguistic Inquiry and Word Count (LIWC) is a text analysis tool developed by James W. Pennebaker and King [13]. It estimates to what level people use different categories of words in large group texts, such as emails, essays or poems. LIWC can also find out the degree of positive or negative emotions, causal words, self-references and 70 other language dimensions used in any text. Hundreds of Microsoft Word documents or standard ASCII text files can be analysed using LIWC program in seconds. You can also build dictionaries of your own to study dimensions of language pertaining to your interests by utilizing LIWC2007.

b) MRC Psycholinguistic Database:

The MRC Psycholinguistic is a dictionary comprising of 150837 words with 26 linguistic and psycholinguistic attributes [12]. It may be applied to psychology or linguistics to formulate sets of experimental inputs, or in computer science or artificial intelligence for linguistic and psychological descriptions of words.

c) SenticNet :

SenticNet 3 is most widely used lexical resource consisting of 30,000 concepts plus their polarity scores ranging from -1.0 to +1.0. The beta version of SenticNet 3.0 comprises of 13,741 concepts, out of which 7626 are multi-word expressions, e.g. high pay joy. SenticNet has 6452 concepts which already exist in WordNet 3.0 where as 7289 of concepts does not. Most of the remaining 7289 concepts are multi-word concepts like make mistake, apart from 82 single-word concepts like telemarketer or against.

d) ConceptNet:

ConceptNet is a semantic network that is specifically used for interpreting text written by individuals. A typical concept network is made up of nodes and links labeled with relationships between them. Each node represents a word or

small phrase of natural language. The nodes are referred to "concepts" or "terms". These relationships are beneficial for searching particular information, answering questions, and understanding goals of individuals. For example, consider the fig. 1, given the two concepts *person* and *cook*, an assertion between them is Capable Of, i.e. *a person is capable of cooking*

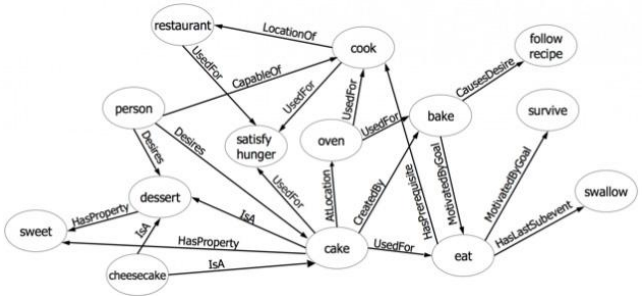

Fig.1: An example for Concept network [17]

e) The EmoSenticNet:

EmoSenticNet dataset [19] is a lexical resource that assigns six WordNet Affect (WNA) emotion labels to SenticNet concept. It contains about 5700 common-sense knowledge concepts, including those concepts that already exist in the WNA list, plus their affective labels in the set {sadness, surprise, anger, joy, disgust, fear}. This resource is useful for sentiment analysis, opinion mining, sentiment polarity detection, social network analysis, emotion analysis, etc.

f) EmoSenticSpace:

To develop a desirable knowledge base for emotive reasoning the authors of [22] employed ‘‘blending’’ technique on EmoSenticNet and ConceptNet. Blending carries out inference over multiple data sources simultaneously, taking advantage of the overlap between them [20]. Essentially, two sparse matrices are combined into a single matrix linearly; sharing the information between the two initial sources. EmoSenticNet is represented as a directed graph like ConceptNet before blending. For example, the concept party is attributed the emotion joy. Then, these concepts are represented as two nodes and assertion HasProperty is added on the edge directed from the node party to the node joy. Then, these graphs are converted to sparse matrices to perform blending. Later Truncated Singular Value Decomposition (TSVD) is performed on the resultant matrix to delete those components constituting comparatively low variations in the data. After discarding only 100 components of the blended matrix are kept to obtain a good approximation of the original matrix. The resulting 100-dimensional space is clustered by means of sentic medoids [21].

2) Classification Methods:

Different classification techniques used to train the classifiers for prediction are as follows

a) Decision Tree (DT):

A decision tree is a tree like structure, consisting of several nodes. The topmost node is called the root node. Each node (internal) shows a test on an attribute, a branch denotes a test outcome, and a leaf node maintains a class label. Different attribute selection measures are used while constructing a tree in order to select the attribute which can best partition the tuples into distinct classes. Many branches from the decision tree may reflect outliers or noise in the training dataset. Tree pruning tries to recognize and discard such branches, to improve the accuracy of classification.

b) C4.5:

The C4.5 is a based on decision tree algorithm. It employs a divide-and-conquer approach for constructing the decision tree. C4.5 applies information entropy concept to construct a decision trees from training dataset. The training dataset s = $s_{1,}$ $s_{2,..}$ is already classified samples. Each sample s_i consists of a p-dimensional vector ($x_{1,i}$,$x_{2,I}$,……$x_{p,i}$) where x_j represent sample’s attributes or features, in addition to the class in which s_i appears.

At each of the node, C4.5 selects the attribute select the attribute which can best partition the set of samples into subsets enriched in one or the other class. The criterion used for splitting is the normalized information gain. The attribute having highest normalized information gain is selected for decision making. The C4.5 algorithm is then repeated on the smaller subsets.

The C4.5 algorithm has following base cases.

- All the samples that are part of the list belong to the same class. In this case, it just creates a leaf node stating to select that class.
- Not any of the attributes provide any information gain. When this happens, it creates a decision node higher up the tree utilizing the expected value of the class.
- In case previously unseen class found. In this situation it creates a decision node higher up the tree utilizing the expected value.

c) k-Nearest Neighbor (k-NN)

This algorithm is based on learning. It achieves this by comparing a given test tuple with training tuples. Training tuples are represented by *n* attributes. Each of the tuple makes up a point in an *n*-dimensional space. Thus all the training tuples are stored in an *n*-dimensional pattern space. In case unknown tuple is given, k-nearest neighbor (k-NN) classifier looks into the pattern space to find *k* training tuples closest to the unknown tuple. These are the *k*-nearest neighbors of the unknown tuple. Euclidean distance is used to define “Closeness”.

d) Linear Regression:

The continuous valued functions are modeled using linear regression. It is widely used because of its simplicity. Generalized linear models present a theoretical foundation to LR for modeling categorical response variables. Poisson regression and Logistic regression are common generalized linear models. Logistic regression models the probability of event taking place as a linear function of a set of predictor variables. Poisson regression is commonly applied to count data as it often displays a Poisson distribution.

e) Naïve Bayes (NB):

These classifiers are statistical classifiers and are commonly used for machine learning. Given a document, NB utilizes the joint probabilities of words and categories to calculate the probabilities of categories. The naïve part of NB is an assumption of word independence, i.e. given a category;

it is assumed that the conditional probability of a word is independent from the conditional probabilities of other words given in that category. Due to this assumption the calculation of the NB classifiers becomes a lot more effective than non-naïve Bayes approaches since it does not utilize word combinations as predictors.

f) Support Vector Machine (SVM):

In a support vector machine algorithm, nonlinear mapping is used for the transformation of the actual training data into a higher dimension. SVM searches for the linear optimal separating in this new dimension. The hyperplane which is a "decision boundary" is responsible for separating the tuples of one class from another. Hyperplane can always separate data from the two classes with a suitable nonlinear mapping. Support vectors and margins are used to find this hyperplane. It is possible that the fastest SVM can be very slow, but because of their capability of modeling complex nonlinear decision boundaries they are highly accurate. They are much less subjected to over fitting.

g) Linear Classifier:

In machine learning, most of statistical classifiers attempt is to utilize the object's characteristics to recognize which class it belongs to. To achieve this linear classifier makes the decision associated with classification on the basis of linear combination value of the characteristics. An object's characteristics referred to as feature values are generally given to the machine in a vector called a feature vector.

h) Multilayer Perceptron (MLP) neural network:

The multi layer perceptron is most commonly used neural network model. Since it requires the knowledge of expected output in order to learn such a neural network is known as a supervised network. This type of network is essential for creating a model that can accurately map the input to the desired output making use of the historical data. This model can then be used to obtain the output when the expected output is not known. The MLP uses an algorithm called backpropagation for learning. The idea behind backpropagation is: the neural network is repeated provided with input data. Then an error is computed by comparing the output of neural network with expected output. This error is then back propagated (fed back) to the neural network and employed to adjust the weights so that, with each iteration there error reduces and the neural model gets closer to producing the expected output. This is known as "training".

i) Ensemble of classifiers :

An ensemble classifier combines the decisions of the individual classifiers in order to enhance the accuracy final decision. Combination of a several trained classifiers yields a performance that is greater than any single classifier would produce since errors generated by one classifier may be corrected by the other. The combination of SVM and k-NN ensemble classifier has an excellent performance on various datasets. Support vector machine classifier is utilized in classification phase, with different kernels: Radial Basis Kernel, Linear and Polynomial kernel. The performance of classification of support vector machine and Naive Bayes of is compared with that of SVM-KNN classifier. The results depict that SVM-KNN model has better classification accuracy than the other.

III. RELATED WORK

Prediction refers to classification of unknown data or to forecast trends .Predicting categorical values is referred to as classification, but if the goal is to model values or continuous functions it is referred to as estimation [11]. Different machine leaning prediction techniques are used for mining social media data. Machine learning includes three strategies: supervised, unsupervised or semi-supervised.

In supervised learning the system is provided with a set of labeled (pre-classified) data, called the training set, to train the predictor. The classifier then classifies new data using the pre-classified data. In unsupervised learning the system is just provided with unlabelled data. The system learns by producing different patterns of what it is exposed to. Semi-supervised learning is in-between supervised and unsupervised learning, which means both the labeled and unlabelled data are used to train the classifier.

Personality prediction involves determining personality traits based on the Big Five Model which is currently the most popular one [12].Various machine learning algorithms mentioned in previous section can be used in this task.

The study [12] was the first one that attempted to relate social media profiles and personality traits. The authors first created a Twitter form consisting of a Big Five Personality Inventory containing 45 questions. Each user was evaluated based on their inventory and their 2000 most recent Tweets. MRC Psycholinguistic Database and Linguistic Inquiry and Word Count (LIWC) were used to extract linguistic information from their messages. Then the extracted linguistic information and the results of the personality tests were then input into a correlation table, and finally, Gaussian and ZeroR processes were used to predict personality based on the Big Five.

In [2] to associate personality scores to Twitter users, they gathered data from a Facebook application called myPersonality. myPersonality users can give their consent to share their personality scores and profile information, and around 40% of them choose to do so. They consider all users who specified their twitter accounts on their Facebook profiles, verified the matching between Facebook and Twitter accounts, and end up having 335 Twitter users. They performed the Big Five personality test on those users. They studied the relationship between the personality traits of the Big Five Model and five types of micro blog users: listeners (those who follow many users); popular (those who are followed by many users); highly read (those who are often 'listened to' in other playlists); and two types of influence indices (TIME and Klout). Using these, the authors created a correlation table and then performed regression by the M5 Rules algorithm to predict personality of profiles.

[14] made use of demographic and text-based attributes extracted from Facebook profiles to predict personality. This study used 537 Facebook profiles and each user was asked to answer a 45-question in order to identify personality based on the Big Five personality index. Then a set of attributes such as age, gender, location, length of biography and quotes, relationship status, and the number of friends, photos, interests, and comments provided, where extracted in order

to define each individual. Then using these predictions individuals were ranked in terms of the five traits, identifying which users would appear above or below 5% or 10% of each trait. They employed numeric prediction models including linear regression, REPTree, and decision tables. Their results showed that it is possible to find the top 10% most open individuals with almost 75% accuracy, and across all traits it predicted the top 10% with at least 34.5% accuracy. The authors explained that these results have privacy implications as they allow advertisers to concentrate on a specific subset of users based on their personality traits. In the same year, a similar study was performed using 2916 Twitter profiles [5].

The study by [3] focused on predicting the dark triad personality (Narcissism, Machiavellianism, and psychopathy) in social media using machine learning. The authors evaluated the predictive ability of NB, SVM, C4.5 and Random Forests in 2927 Twitter profiles from 89 countries and recognized significant correlations among the Twitter users and dark triads. Self-reported ratings were formulated from the Short Dark Triad (SD3) questionnaire supplying the measures of psychopathy, Machiavellianism and narcissism; and Ten Item Personality Inventory (TIPI), supplying measures of agreeableness, openness, conscientiousness, extraversion and neuroticism to extract personality for each user. 3200 Tweets were downloaded using Twitter API and then analysed using Linguistic Inquiry and Word Count (LIWC). The final result consisted of 586 features such as number of Tweets, number of followers, number of friends, and the frequency of predefined words for each individual. The personal information was removed and a subset of 337 features to be used by the machine learning predictor. The study showed that psychopaths and Machiavellians tend to use more swears words and words associated with anger. The use of Mean Root Mean Square Error (RMSE) and Average Error (MAE) evaluation methods improved the accuracy of their results.

In [17] the authors proposed a new architecture to identify personality making use of the common sense knowledge along with associated affective labels and sentiment polarity. They used essays dataset which contains 2400 essays labeled manually with personality scores for five different personality traits. They extracted several features from the text using LIWC, MRC and combine them with the common sense knowledge based features extracted by sentic computing techniques (SenticNet, ConceptNet, EmoSenticNet and EmoSenticSpace). In particular, they combined common sense knowledge based features with frequency based features and psycho-linguistic features and then used these features in supervised classifiers. For each of the five personality traits an SMO based supervised classifier was designed. The common sense knowledge with sentiment information and affective labels increases the accuracy of the frameworks which only use frequency based analysis and psycho-linguistic features at lexical level.

The study [15] highlights drawbacks of supervised machine learning i.e. limited availability and high cost of obtaining training (labeled) data, and thus provides a solution based on ensemble learning. In this approach classifiers were constructed using of information from datasets of different genres, personality classification systems and even different languages. Five meta-learning experiments were carried out with Facebook data, one for each personality trait. The data included anonymous authors, status updates in text and a number of social network measures. As attributes they used the 2000 most frequent character trigrams. In each of the experiments performed, the ensemble (meta) learner used a 10-fold cross validation.

In [16] the authors proposed a new architecture called PERSOMA. They first obtained 18,435 Tweets (sum of the three datasets) and then clustered them into the 41 groups. In the preprocessing stage meta-attributes were extracted from tweets and a metabase was created. The metabase was then sent to the transformation module where the multi-label classification problem was transformed into five binary classification problems. The multi-label classification is performed by five classification algorithms, each one responsible for one single class, i.e., personality trait. Three classification algorithms were used namely Naïve Bayes (NB), a Support Vector Machine (SVM), and a Multilayer Perceptron (MLP) Neural Network were used as classifiers. The classifiers were trained using semi-supervised learning, so that the training set increases as new classifications are made, in a transductive semi-supervised learning style. A k-fold cross-validation was used in the semi-supervised learning with k = 4; a single fold for training and three for testing. In order to form training set the (small number) tweets previously classified using PRec. Within every new classified fold its labeled objects were added to the training set.

IV. DISCUSSION

Table II summarizes the approaches reviewed in the previous section along with their strengths and limitations. It can be inferred that most of the approaches typically require the users to fill a form containing several questions and then use this inventory to predict personality based on big five model. Some of them also consider the user's profile data for prediction. Social media-oriented people tend to publish a lot about themselves by status updates, self-description, interests and photos. But they do not share details they find sensitive. They keep it private either to themselves or make it available only to a certain group of people. Some users deliberately fake their personal information such as Birth date, location, and work , status and/or even create a fake identity just to become influential e.g. to increase number of follows; get more likes etc. They decide what must be shared and what not on their convenience. As a result the social media data available may either be fake, missing or cannot be accessed due to privacy issues. Thus results of personality prediction cannot be accurate. Most of the systems simply work with a single line of text. They extract grammatical information such as number of words, number of positive and negative words etc. These attributes are then used in further stages of prediction. They do not consider social behaviour information such as number of friends/followers, number of tweets, number of has tags etc. These attributes may help in determining for e.g. how frequently the person makes us of social media for interaction.

TABLE II: SUMMARY OF PERSONALITY PREDICTION APPROACHES

Name of the paper	Techniques	Strengths	Limitations
Predicting Personality From Twitter [12]	1) *Data collection:* Twitter application form and publicly available profile data; 2) *Information Extraction:* LIWC Tool and MRC Psycholinguistic Database; 3) *Analysis:* Pearson correlation; *4) Machine Learning:* Gaussian Process & ZeroR.	Can make a prediction for each of the five personality factors between 11% - 18% of the actual values.	1) Some language features were not considered during analysis e.g. misspelled words on Twitter 2) It yielded less impressive results for conscientiousness, extroversion & neuroticism 3) Personality scores between friends were overlooked.
Our Twitter Profiles, Our Selves: Predicting Personality with Twitter [2]	1) *Data collection:* Facebook application - myPersonality and publicly available profile data; 2) Considered all users who specified their twitter accounts on their Facebook profiles; 3) *Analysis:* Pearson product-moment correlation; 4) *Machine Learning:* M5 Rules algorithm.	The myPersonality app provided a high test result and its users gave their consent to share their personality score and profile information. Thus using the three count (following, followers, and listed counts) they could predict user's personality better.	1) Prediction becomes difficult when people create fake accounts, on fake some information and even when relevant information is not available for analysis. 2) The prediction of traits is made informally based on intuitions and thus they cannot guarantee the level of accuracy.
Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets [3]	1) *Data collection:* Twitter application form and publicly available data; 2) *Information Extraction:* LIWC Tool; 3) *Analysis:* zero-order Spearman's correlation; 4) *Machine Learning:* C4.5, Naïve Bayes, Random Forests and Support Vector Machines.	It successfully proved that there are relationships between Big Five traits, Dark Triad and Twitter activity. They made use of Root Mean Square Error (RMSE) and Mean Average Error (MAE) for evaluation which enhanced the accuracy of their results.	1) Using Twitter alone is likely to be both insufficient for personality prediction and also error prone. 2) Due to the maximum limit of 140 characters, tweets may be written in informal language. Using LIWC dictionary alone to analyze the data is not sufficient. It also results in a very small number of feature extractions. 3) The study was purely based on self assessment questionnaires, which could be easily manipulated by people in order to induce a measurement error.
Machine prediction of personality from Facebook profiles [14]	1) *Data collection:* From survey consisting of 45-questions and profile information. 2) *Information Extraction:* LIWC Tool; 3) *Machine Learning:* Linear regression, REPTree, and decision tables	By using certain prediction models, they could identify the topmost 10 % of Open individuals with almost 75% accuracy, and also predict the top 10% of individuals across all traits and directions, with at least 34.5% accuracy. These results would allow marketers and other interested parties to focus on specific subsets of users based on their profile information and create advertising more closely tailored to those users.	1) The attackers seeking to perform social engineering attacks could determine which subset of the population is most susceptible. 2) The performance of prediction models will degrade with the increase in number of individuals.
Common Sense Knowledge Based Personality Recognition from Text [17]	1) *Data collection:* Essays dataset containing 2400 essays labeled manually with personality scores 2) *Information Extraction:* LIWC, MRC database, SenticNet, EmoSenticNet, EmoSenticSpace, ConceptNet ; 3) *Machine Learning:* SMO (Sequential Minimal Optimization) model.	The common sense knowledge with sentiment information and affective labels increased the accuracy of the existing frameworks which only use frequency based analysis and psycho-linguistic features at lexical level.	1) Agreeableness is most difficult trait to identify among all traits.
Ensemble Methods for Personality Recognition [15]	1) *Data collection:* Facebook Data (test) and essays(training); 2) *Machine Learning:* SMO [ensemble (meta) learner]	Ensemble methods successful improved the accuracy of systems by combining the predictions of different component classifiers.	1) The personality recognition of essay data, using output of classifiers trained on the Facebook data as part of the out of- genre ensemble caused performance deteriorated for other traits.

A multi-label, semi-supervised classification approach applied to personality prediction in social media [16]	1) *Data collection:* Twitter dataset 2) *Information Extraction:* LIWC Tool and MRC Psycholinguistic Database; 3) *Machine Learning:* Naïve Bayes, Support Vector Machines, Multilayer Perceptron Neural Network.	It works with group of texts, rather than a single text, and does not rely on users' profiles and has an accuracy of 83% for some traits.	1) Tweets may be written in slang language and contain special characters. Therefore, the automatic analysis of Twitter message is difficulty. 2) Openness and conscientiousness were the most difficult trait to predict. This may be because semi-supervised learning approach used is based on grammar, and does not take social behaviors in account.

Thus this study provides following insights: For information extraction the system can use approaches such as LIWC, MRC database, SenticNet, EmoSenticNet, EmoSenticSpace, ConceptNet together so that more number of features can be extracted. To determine the personality accurately, the prediction system must utilize both the grammatical and social behavior, as well as work with a group of tweets/posts. The classification should make use of ensemble of classifiers into to improve the accuracy of prediction. Finally using twitter alone is insufficient to predict personality, thus other social networking sites must also be considered to improve the accuracy of personality identification.

V. CONCLUSION

Social media provides a platform where the users can share information as well as get feedbacks from colleagues and friends. The exposure of views by individuals encourages other users to comment and share their ideas as well as information about themselves. This reveals the personality which may be useful in many areas such as marketing, business intelligence, and psychology. Identification of personality traits helps us to understand user behavior and trends. The Big Five model allows the identification of traits using linguistic information. This survey paper provided a brief review of the techniques that have been used for personality prediction and discussed some strengths and limitations of these approaches. Thus it concluded that in order to improve prediction, it is necessary to consider both group of texts as well as social behavioral aspects of a user on multiple social media (e.g. Twitter, Facebook, and LinkedIn).

REFERENCES

[1] G. Barbier & H. Liu, "Data mining in social media" in C. C. Aggarwal (Ed.), Social network data analytics, pp. 327–352, US: Springer, 2011.
[2] D. Quercia, M. Kosinski, D. Stillwell, & J. Crowcroft, "Our twitter profiles, our selves: predicting personality with twitter", in IEEE international conference on privacy, security, risk, and trust, and IEEE international conference on social computing, pp. 180–185, 2011.
[3] C. Sumner, A. Byers, R. Boochever, & G.J. Park, "Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets", in 11th international conference on machine learning and applications, pp. 386–393, 2012.
[4] D. Garcia, & S. Sikström, "The dark side of Facebook: semantic representations of status updates predict the dark triad of personality", in Personality and Individual Differences, 2013.
[5] R. Wald, T. M. Khoshgoftaar, A. Napolitano, C. Sumner, "Using twitter content to predict psychopathy", in Proceedings of the 2012 11th international conference on machine learning and applications—Volume 02, pp. 394–401, Washington, DC, USA.
[6] H. P. Martinez, Y. Bengio & G. Yannakakis, "Learning deep physiological models of affect", in IEEE Computational Intelligence Magazine, 8(2), 20–33, 2013.
[7] E. Cambria, B. Schuller, B. Liu, H. Wang & C. Havasi "Knowledge-based approaches to concept-level sentiment analysis", IEEE Intelligent Systems, 28(2), pp.12–14, 2013.
[8] E. Tupes and R. Christal, "Recurrent personality factors based on trait ratings", Journal of Personality, vol. 60, no. 2, pp. 225–251, 1992.
[9] R. McCrae and O. John, "An introduction to the five-factor model and its applications", Journal of personality, vol. 60, no. 2, pp. 175–215, 1992.
[10] J. Digman, "Personality structure: Emergence of the five-factor model", Annual review of psychology, vol. 41, no. 1, pp. 417–440, 1990.
[11] J. Han, M. Kamber, & J. Pei, "Data mining: concepts and techniques", (3rded.), Morgan Kaufmann, 2011.
[12] J. Golbeck, C. Robles, M. Edmondson & K. Turner, "Predicting personality from twitter", in IEEE international conference on privacy, security, risk and trust, and IEEE international conference on social computing , pp. 149–156, 2011.
[13] J. W. Pennebaker, & L. A. King, "Linguistic styles: language use as an Individual difference", Journal of Personality and Social Psychology, vol. 77, no. 6, pp. 1296–1312, 1999.
[14] R. Wald, T. Khoshgoftaar & C. Sumner, "Machine prediction of personality from Facebook profiles", in 2012 IEEE 13th international conference on Information Reuse and Integration (IRI), pp. 109–115.
[15] B. Verhoeven, W. Daelemans, & T. De Smedt, "Ensemble methods for personality recognition", in Proceedings of the workshop on computational personality recognition, 2013.
[16] Ana Carolina E.S. Lima, Leandro Nunes de Castro, "A multi-label, semi-supervised classification approach applied to personality Prediction in social media" in Affective Neural Networks and Cognitive Learning Systems for Big Data Analysis, vol. 58, October 2014, pp. 122–130.
[17] S. Poria, A. Gelbukh, B, Agarwal, E. Cambria & N. Howard, "Common sense knowledge based personality recognition from text", in F. Castro, A. F. Gelbukh, & M. González (Eds.), Lecture Notes in Computer Science: vol. 8266, Advances in Soft Computing and Its Applications, MICAI(2), pp. 484–496, Springer.
[18] O. P. John & S. Srivastava, "The big-five trait taxonomy: history, measurement, and theoretical perspectives", (2nd ed.), New York: Guilford Press, 2001.
[19] S. Poria, A. Gelbukh, A. Hussain, D. Das, S. Bandyopadhyay , " Enhanced SenticNet with Affective Labels for Concept-based Opinion Mining" , IEEE Intelligent Systems, vol. 28, issue 2, 2013, pp 31–38.
[20] C. Havasi, R. Speer & J. Pustejovsky, "Automatically suggesting semantic structure for a generative Lexicon ontology", in Generative Lexicon, 2009.
[21] E. Cambria, N. Howard, J. Hsu & A. Hussain, "Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics" In IEEE SSCI, pp. 108–117, 2013.
[22] S. Poria, E. Cambria, A. Hussain , Guang-Bin Huang, "Towards an intelligent framework for multimodal affective data analysis" in Neural Networks, Vol. 63, March 2015, pp. 104–116, (2014)