International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

A SURVEY ON PERSONALITY PREDICTION USING DIGITAL FOOTPRINTS IN SOCIAL MEDIA

M.Suresh[1] D.Nagendrababu[2],S.Nakkiran[3], K.Vanjinathan[4]

[1]Assitant professor, Department of Information Technology, mail4sureshuni@gmail.com.
[2, 3, 4] B.TECH, 4th year student, Department of Information Technology, Manakula Vinayagar Institute of Technology, Pondicherry, India.
[2]nagfury@gmail.com,[3]nakkiran93@gmail.com,[4]vanjinathan03@gmail.com.

Abstract - *Big data is now a popular term used to describe the exponential growth and availability of data, both structured and unstructured data. It vary from conventional cloud services, one of the main features of big data services is coupled between data and computation. This can be conducted only when the related data is available. The social networking sites like Twitter, Face book, LinkedIn and YouTube allow the users to create and share content related to different subjects, reflect their activities, feelings, thoughts and opinions. This data provides the information about human behavior and social interactions. It makes it possible to understand the user's interests and their needs. This information may be used to survey the consumer's opinions in conceptualizing different business strategy. This paper reviews the techniques used in analyzing social media data to identify important of personality traits. The personality trait is the characteristics and quality of a person. It can be used in different areas such as psychology, marketing and sociology. A parallelism among individual's personality traits and linguistic information are processed for analytics.*

Key Words: **Big data; Big Five model; Psychology; Lexical; Resources; Personality; Social Media.**

1. INTRODUCTION

Social networking on the web has become an essential component of everyday life. It has radically changed the ways in which people convey their opinions and feelings. Social networking sites such as Facebook, Twitter and YouTube are based on human interaction which is 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. 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 is extent in volume and reveals different aspects of human behaviour and social interaction. 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. This serves as a key factor for marketers who want to create a fixed image in the minds of the customers for their product, brand, or organization and therefore identify which products to recommend to the user. Our technique leverages what people convey in social media to find distinctive words, phrases, and topics as functions of known his/her characteristics and aspects that others can see. Thus, the analysis of social media data makes it possible to identify important personality traits, such as characteristics or qualities which describes his/her personality. 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.

2. TECHNIQUES

2.1 Data analysis and Classification Methods

This section first contains the various lexical resources used in analyze the text messages and then explains classification methods.

2.1.1 LIWC (Linguistic Inquiry and Word Count)

Linguistic Inquiry and Word Count (LIWC) is a text analyzing tool developed by James W. Pennebaker and King. 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.

2.1.2 MRC Psycholinguistic Database

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

sychological descriptions of words.

2.1.3 SenticNet

SenticNet 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 whereas 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. For an example a sentence like "I am going to the market to buy vegetables and some fruits" this parser extracts concepts such as 'go_buy','go_to_market', 'market, buy_fruit', and 'some_fruits'.

2.1.4 EmoSenticSpace

To develop a desirable knowledge base for emotive reasoning the authors of employed blending technique on Emo Sentic Net and ConceptNet. Blending carries out inference over multiple data sources simultaneously, taking advantage of the overlap between them. 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 senticmedoids.

2.2 Classification Methods

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

2.2.1 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.

2.2.2 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.

2.2.3 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 hyper-plane which is a decision boundary is responsible for separating the tuples of one class from another. Hyper-plane can always separate data from the two classes with a suitable nonlinear mapping. Support vectors and margins are used to find this hyper-plane. It is possible that the fastest SVM can be very slow, but because of their capability of modelling complex nonlinear decision boundaries they are highly accurate. They are much less subjected to over fitting.

2.2.4 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.

2.3 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, behaviour, temperament and the mind defines a 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, 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. Similarly for determining personality. Personality can vary depending on different situations. In analysis of the personality structure, definition of the Big Five Model or Five Factor Model came into use. The Big Five traits can be described as follows:

- **Openness** is related to imagination, creativity, curiosity, tolerance, political liberalism, and appreciation for culture. People scoring high on openness like change, appreciate new and unusual ideas, and have a good sense of aesthetics.

- **Conscientiousness** measures the preference for an organized approach to life in contrast to a spontaneous one. Conscientious people are more likely to be well organized, reliable, and consistent. They enjoy planning, seek achievements, and pursue long-term goals. Non conscientious individuals are generally more easy going, spontaneous, and creative. They tend to be more tolerant and less bound by rules and plans.

- **Extroversion** measures a tendency to seek stimulation in the external world, the company of others, and to express positive emotions. Extroverts tend to be more outgoing, friendly, and socially active. They are usually energetic and talkative; they do not mind being at the centre of attention and make new friends more easily. Introverts are more likely to be solitary or reserved and seek environments characterized by lower levels of external stimulation.

- **Agreeableness** relates to a focus on maintaining positive social relations, being friendly, compassionate, and cooperative. Agreeable people tend to trust others and adapt to their needs. Disagreeable people are more focused on themselves, less likely to compromise, and may be less gullible. They also tend to be less bound by social expectations and conventions and are more assertive.

- **Emotional stability** (opposite referred to as neuroticism) measures the tendency to experience mood swings and emotions, such as guilt, anger, anxiety, and depression. Emotionally un- stable (neurotic) people are more likely to experience stress and nervousness, whereas emotionally stable people (low neuroticism) tend to be calmer and self-confident. Research has shown that personality is correlated with many aspects of life, including job success, attractive- ness, marital satisfaction, infidelity, and happiness. The main limitations of classical personality studies are, however, the size of the samples, often too poor for statistical validation, and their strong bias toward white, educated.

Table I: Big Five Model

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

2.LITERATURE REVIEW

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. Different machine leaning prediction techniques are used for mining social media data. Machine learning includes three strategies: supervised, unsupervised or semi-supervised. In 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 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 and highly read and two types of influence indice. Using these, the authors created a correlation table and then performed regression by the M5 Rules algorithm to predict personality of profiles.

3.1 Predicting Personality from Twitter

Social media is a place where users present themselves to the world, revealing personal details and insights into their lives. To understand how some of this information can be utilized to improve the users' experiences with interfaces and with one an-other it has been shown to be useful in predicting job satisfaction, professional and romantic relationship success, and even preference for different interfaces. They begin to validate this metric by showing that positive and negative word use in status updates covaries with self-reported satisfaction with life, and also note that the graph shows peaks and valleys on days that are culturally and emotionally significant. 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.[2]

3.2 Predicting Personality on Social Media with Semi-Supervised Learning

Personality research on social media is a hot topic recently due to the rapid development of social media as well as the central importance of personality study in psychology, but most studies are conducted on inadequate label samples. The local linear semi-supervised regression algorithm has been employed to establish prediction model. People are happier on weekends, but the morning peak in positive affect is delayed by 2 hours, which suggests that people awaken later on weekends. 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, 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 behaviours in account.[3]

3.3 What you're Face Vlogs About: Expressions of Emotion and Big-Five Traits Impressions in Youtube

Social video sites where people share their opinions and feelings are increasing in popularity. The face is known to reveal important aspects of human psychological traits, so the understanding of how facial expressions relate to personal constructs is a relevant problem in social media. They use the Computer Expression Recognition Toolbox (CERT) system to characterize users of conversational vlogs. The cue sets are first used in a correlation analysis to assess the relevance of each facial expression of emotion with respect to Big-Five impressions obtained from crowd-observers watching vlogs. 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.[4]

3.4 Our Twitter Profiles, Our Selves: Predicting Personality with Twitter

Psychological personality has been shown to affect a variety of aspects: preferences for interaction styles in the digital world and for music genres, for example the design of personalized user interfaces and music recommender systems might benefit from understanding the relationship between personality and use of social media. They then show a way of accurately predicting a user's personality simply based on three counts publicly available on profiles: following, followers, and listed counts. Knowing these three quantities about an active user, one can predict the user's five personality traits with a root-mean-squared error below 0.88 on a scale. They analysed the use of emotion words for approximately 100 million Facebook users since September of 2007. "Gross national happiness" is operationalized as a standardized difference between the use of positive and negative words, aggregated across days, and present a graph of this metric. They discuss the development and computation of this metric, argue that this metric and graph serves as a representation of the overall emotional health of the nation, and discuss the importance of tracking such metrics.[6]

3.5 Data Set: Facebook Status Updates

Their complete dataset consists of approximately 19 million Face book status updates written by 136,000 participants. Participants volunteered to share their status updates as part of the My Personality application, where they also took a variety of questionnaires. They restrict our analysis to

those Facebook users meeting certain criteria: They must indicate English as a primary language, have written at least 1,000 words in their status updates, be less than 65 years, and indicate both gender and age. This resulted in N~74,941 volunteers, writing a total of 309 million across 15.4 million status updates. From this sample each person wrote an average of 4,129 words over 206 status updates, and thus 20 words per update. Depending on the target variable, this number slightly varies as indicated in the caption of each result. The personality scores are based on the International Personality Item Pool proxy for the NEO Personality Inventory Revised. Participants could take 20 to 100 item versions of the questionnaire, with a retest reliability of aw0:80.. The statuses ranged over 34 months, from January 2009 through October 2011. Previously, profile information (i.e. network metrics, relationship status) from users in this dataset have been linked with personality, but this is the first use of its status updates.[7]

3. DISCUSSION

Table II summarizes all the approaches reviewed in the previous section along with their advantages and limitations.

The Digital footprints are the traces left by the user in social network suck as search contents, updated status, comments. These footprints can be collected and can be used on different purposes. The concept of predicting personality from social media fully depends upon these footprints. The LIWC is tool is used correlate collected data and predicts user personality. The SEMI-SUPERVISED LEARNING is another technique which uses local semi supervised regression algorithm to establish the prediction model, this approach is based on the grammar and does not take social behaviours in account. Social video streaming sites like YouTube uses the CERT(Computer Expression Recognition Toolbox) to characterize the user of conversational volgs. The Big-Five Impression obtained from observers watching volgs. These are some methods to collect the reliable information from the digital footprints which are used to predict the user's psychological characteristics and behaviour. They extract grammatical information such as number of words, number of positive and negative words etc. These attributes are used in further stages of prediction. They do not consider social behaviour information such as number of friends or followers, number of tweets, number of has tags etc.

Table II: Summary Of Personality Prediction Techniques

Name of the paper	Year	Techniques	Advantages	Limitations
Predicting Personality from Twitter [2]	2011	Information Extraction: LIWC Tool and MRC Psycholinguistic Database.	Can make a prediction for each of the five personality factors between 11%-18% of the actual values.	Some language features were not considered during analysis. e.g. misspelled words on Twitter.
Predicting Personality on Social Media With Semi-Supervised Learning [3]	2008	Data collection: Facebook app - myPersonality and publicly available profile data;	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.	Tweets may be written in slang language and contain special characters. Therefore, automatic analysis of Twitter message is difficulty.
What You're Face Vlogs About: Expressions of Emotion and Big-Five Traits Impressions in Youtube [4]	2014	Computer Expression Recognition Toolbox (CERT) system to characterize users of conversational vlogs	The common sense knowledge with sentiment information and affective labels increased the accuracy of the existing frameworks.	Using CERT dictionary alone to analyze the data is not sufficient. It also results in a very small number of feature extractions.
Our Twitter Profiles,Our Selves:Predicting Personality with Twitter [6]	2011	Considered all users who specified their twitter accounts on their Facebook profiles. *Analysis:* Pearson product-moment correlation.	This app is 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.	Prediction becomes difficult when people create fake accounts, on fake some information. The prediction of traits is made informally based on intuitions and thus they cannot guarantee the level of accuracy.
Data Set: Facebook Status Updates [7]	2014	Information Extraction: LIWC, MRC database, SenticNet, EmoSenticSpace, ConceptNet	The common sense knowledge with sentiment information and affective labels increased the accuracy.	Agreeableness is most difficult trait to identify among all traits.

3. CONCLUSION

Online social media are particularly promising resource for the study of people, as updates are self-descriptive, personal, and have emotional content. Language used is objective and quantifiable behavioural data, and unlike surveys and questionnaires, Social media language allows researchers to observe individuals as they freely present themselves in their own words. Differential language analysis (DLA) in social media is an unobtrusive and non-reactive window into Personality, Gender, Age in Social Media Language the social and psychological characteristics of people's everyday concerns. Most studies linking language with psychological variables rely on a priori fixed sets of words, such as the LIWC categories the data. Here we show the benefits of an open-vocabulary approach in which the words analysed are based on the data itself. The extracted words, phrases, and topics from millions of Facebook messages and found the language that correlates most with gender, age, and five factors of personality. Results have face validity, tie in with other research, suggest new hypotheses and give detailed insights. Over the past one-hundred years, surveys and questionnaires have illuminated our understanding of people. We suggest that new multipurpose instruments such as Five Factor model emerging from the field of computational social science shed new light on psychosocial phenomena.

REFERENCES

[1] Renaud Lambiotte,Michal Kosinski, "Tracking the Digital Footprints of Personality", in IEEE rans, 2015, Volume:102, no:12,pp.1934 – 1939.

[2] J. Golbeck, C. Robles, M.Edmondson, and K. Turner,"Predicting personality from Twitter," in Proc. Int. Conf.Privacy Security Risk Trust/IEEE Int. Conf. Social Comput., 2011,pp.149–156.

[3] D. Evans, S. Gosling, and A.Carroll, "Predicting Personality on Social Media With Semi supervised Learning," in Proc.Conf. Weblogs Social Media,2008, pp. 45–50.

[4] J.-I. Biel and D. Gaticaerez,"What You're Face Vlogs About: Expressions of motion and Big-Five Traits Impressions in Youtube", *IEEE, Trans.2014, Multimedia*, pp.41-55.

[5] B. Bi, M. Shokouhi, M. Kosinski, and Graepel, "Finding Waldo:Learning about Users from Their Interactions," in Proc. Int.World-Wide Web Conf., 2013, pp. 131–140.

[6] D. Quercia, M. Kosinski, D.Stillwell, and Crowcroft, "Our Twitter profiles, our selves:Predicting personality with Twitter," in Proc. Int. Conf.Privacy Security Risk Trust/IEEE Int. Conf. Social Comput., 2011,pp.180–185.

[7] S. W. Chew, P. Lucey, S. Lucey, J. M. Saragih, J. F. Cohn, I.atthews and S. Sridharan ," Data Set:Facebook Status Updates" , *IEEE Trans. Syst., Man, Cybern. 2014,B* , vol. 42 , no. 4 , pp.1006 -1016.

[8] L. Backstrom and J. Kleinberg, "Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on Facebook" , in Proc. 17th ACM Conf. Comput.Supported Cooperative Work Social Comput., 2014, pp. 831–841

[9] D. Quercia et al., "Facebook and privacy: The balancing act of personality, gender, relationship currency," in Proc. Conf. Weblogs Social Media, 2012, pp. 306–31.

[10] M. Kosinski, Y. Bachrach, P.Kohli, Stillwell, and T. Graepel, "Manifestations of user personality in website choice and behaviour on online social networks," Mach.Learn., 2013,vol. 95, pp. 357–380,

[11] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," Proc. Nat. Acad. Sci., vol. 110, 2013, pp. 5802–5805,.

[12] H. Schwartz et al., "Personality, gender, age in the language of social media: The open-vocabulary approach," PloS One, vol. 8, no. 9, 2013, pone. 0073791.

[13] Y. Chen, D. Pavlovy, and J. F.Canny, "Large-scale behavioral targeting,' in Proc.Conf. Knowl. Disc. Data Mining, 2009,pp. 209–217.

[14] Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T.iyamoto, "The effects of responsive eye movement and blinking behavior in a communication robot," in IROS, 2006, pp. 4564-4569.

[15] D. Byrne, W. Griffitt, and D. Stefaniak, "Attraction and similarity of personality characteristics," J. Personality Social Psychol., vol. 5, no. 1, pp. 82–90, 1967.

[16] B. W. Roberts, O. S.Chernyshenko, S. Stark, and L. R. Goldberg, "The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires," Personnel Psychol.,2005, vol. 58, no. 1, pp. 103–139.

[17] S. Gosling, S. Gaddis, and S.Vazire, ‘‘Personality mpressions based on facebook profiles,’’ in Proc. Int. Conf. Weblogs Social Media, 2007, vol. 7, pp. 1–4.

[18] Rammstedt and O. John , "Measuring personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory" , *J. Research in Personality* , vol. 41 , pp.203 -212.

[19] A.-L. Vollmer, K. S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. J. Rohlfing, and B. Wredek, "People modify their tutoring behavior in robot-directed interaction for action learning," in DEVLRN '09: Proceedings of the 2009 IEEE 8th International Conference on Development and Learning. Washington, DC, USA: IEEE Computer Society, 2009, pp. 1-6.

[20] F. Broz, I. R. Nourbakhsh, and R. G. Simmons, "Designing pomdp models of socially situated tasks," in Proc of the 20th IEEE International Symposium on Robot and Human Interactive Communication (Ro-Man), 2011.