IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 4.819

Volume 6, Issue 1, June 2021

Personality Prediction Based on Twitter Information

Krupa D. Rajpara[1] and Rupal J. Shilu[2]

Student, Computer Department, Atmiya University, Rajkot, India[1]

Assistant Professor, Computer Department, Atmiya University, Rajkot, India[2]

Abstract: *Personality is an important factor that affects a person's opinions, like-dislike, thoughts, and how a person behaves in different situations. Nowadays social media is a platform where people express their view, how they feel, what they are doing. Every second on average, around 6000 tweets are generated. This data can be analyzed properly to predict the personality of the user. The aim of this paper is to identify a person's personality from the tweets posted by him. Whatever person writes, the word he uses can be used to predict his personality traits. Based on personality we can prioritize advertisement, market plan, job suggestion etc. The aim of this paper is to review personality prediction methods from social media used by different authors around the globe.*

Keywords: Big five personality traits, Data Mining, OCEN, Personality Assessment, Twitter

I. Introduction

Social media is a vital part of people's lives. Use of social media has increased tremendously. The social network's size was 290.5 million monthly active users worldwide, and was projected to keep increasing up to over 340 million users in 2024[1]. Every active user upload text or image according to their choice, personality. A large amount of data is generated every second, which can be analyzed to predict personality.

Traditional method of predicting personality involves a questionnaire. The set of questions are prepared for the user. They answer the question as they would react in a particular scenario. Answers are then analyzed by a professional psychologist and a personality report is made. The problem with the traditional method is it requires the user to answer the question honestly. Also, it requires a lot of time to analyze the user's answer. This method is time consuming and requires trained people to make personality reports. People can also answer a set of questions and then answers are analyzed by advanced machine learning algorithms but here also users need to be honest in the test.

Researchers and professionals are studying various methods and approaches to predict personality from the text written by the user. Text written, word used, number of times a user posts can predict a user's personality traits. There are different models available to predict personality like Big five Model, DISC, Myers-Briggs Type Indicator.

II. Literature Survey

In this section we have discussed work regarding predicting personality from various journals and research papers. Different authors have proposed various approaches to predict personality from written text and different algorithms are used for feature extraction and personality prediction.

S. Arjaria[2] have used MLNB classifier to predict big five personality traits on essay dataset. The algorithm is trained with 10fold cross validation. Essay data set was generated by asking students to write randomly whatever they think for 20 min.

Nadeem Ahmad and Jawaid Siddiquea[3] used DISC assessment. They used Rapidminor tool to extract twitter dataset on which text mining is applied through R language. Psychoanalytic profiling has been a well-received and proven method with scores of techniques that are undertaken to study multi-dimensionality in an individual's personality. Social media has changed the linguistic syntax across the globe and now people seem more comfortable to express themselves with tag words instead of using a statement or complete sentence. These tag words are helpful in generating themes which assists researchers in drawing out respective conclusions with maximum precision.

DOI: 10.48175/568

P. S. Dandannavar[4] provided automatic personality prediction from social media models, which can be used for recruitment, marketing, corporate, counselling, psychological profiling, E-commerce.

Helly. N. Desai and Prof. Rakesh Patel [5] compared supervised algorithm, unsupervised algorithm, semi-supervised algorithm. Digital footprints were used as a dataset.

Sandhya Katiyar, Himdweep Walia, Sanjay Kumar [6] used a support vector machine, Naïve Bayes algorithm to predict personality. Applicants take a survey of 30 questions which is analyzed. Naïve bayes outperformed than support vector machine.

Joel Philip [7] has used machine learning techniques to predict personality from twitter, Facebook, Quora. For Feature extraction authors have used Bag of words, TF-IDF, Word Net. limitation is at a time users will share status on ongoing topics currently happening in the world. In such a situation the system should have information about current trends and topics to better understand the user's personality traits.

Multinomial naïve bayes, Adaboost and LDA algorithm are used by Aditi V.Kunte[8] on twitter dataset. Psychology is a broad domain and personality is an integral part of this. Users independently post their thoughts without being judged. Social media is a promising platform to predict personality.

Decision Tree C4.5 is used by Willy, Erwin B. Setiawan and Fida N. Nugraha[9] along with TF-RF and TF-CHI2 on twitter dataset. SPSS 24 software is used to label questions based on big five inventory. Srilakshmi Bharadwaj [10] identified personality traits based on myers-briggs type indicator using neural network and SVM with LIWC (Linguistic Inquiry and Word Count). Author developed web application gives score for MBTI personality type.

III. Big Five Model

A Big Five model is also known as the Five Factor Model; it focuses on five core factors, known by the acronym OCEAN or CANOE. Gordon Allport and Henry Odbert first formed a list of 4,500 terms relating to personality traits in 1936.The model has received much attention as each of personality traits represents extremely broad category.[11]

Openness to Experience: Imagination, Feelings, actions, ideas

Users who score high in this tend to be curious, have a wide range of interests, and are independent. They are excited to try new things and are adventurous. While people who score less in this trait tend to be predictable, not very imaginative. They are uncomfortable with changes and trying new things so they prefer familiar things.

Conscientiousness: Competence, self- discipline, thoughtfulness, goal-driven

Users who score high on Conscientiousness tend to be hard working, dependable and organized, thoughtful, careful. They have good impulse control which allows them to complete tasks and achieve goals. People who score less on conscientiousness tend to be disorganized, careless, and undisciplined. Hence, they find difficulty in completing tasks and fulfilling goals.

Extraversion: sociability, assertiveness, emotional expression

Users who score high on extraversion enjoy being the center of attention. They are sociable, energized by social interaction, and feel comfortable voicing their opinion. Users who score less on extraversion are known as introverts. They do not like social events, they find it tiring. They are reserved and make less friends.

Agreeableness: cooperative, trustworthy, good-natured

Users who score high in this tend to have straightforwardness, Empathy, Modesty, Compliance, enjoy helping. They are sensitive to other’s needs. Users who score less are demanding, stubborn, unsympathetic. They don’t care about how other people feel.

Neuroticism: tendency toward unstable emotions

Users who score high on neuroticism tend to be more anxious, moody, depressed and shy. they feel insecure and quickly get irritated. Users who scoreless in this trait tend to be calm, composed, secure and self-satisfied.

3.1 Framework for Personality Prediction

Data Extraction: In this phase data is collected from various social media API. An API dataset is freely available for research. This dataset includes 1600000 tweets. It contains six field targets, ids, date, flag, user, text.[12]

Data Preprocessing: After collection of data, raw data needs to be preprocessed. Human language can be understood byusing natural language processing. Preprocessing also requires some steps like case folding (Changing all letters tolowercase), tokenizing (decomposition of tweets into words, removing punctuation mark using unigram method), filtering (selecting important word, removing ‘he’, ‘she’, ‘it’), Stemming, (dropping unnecessary character usually suffix) Term weighting (Assigning weight or value to every word using TF-RF, TF-CHI2) NLTK library function are used for preprocessing.[13]

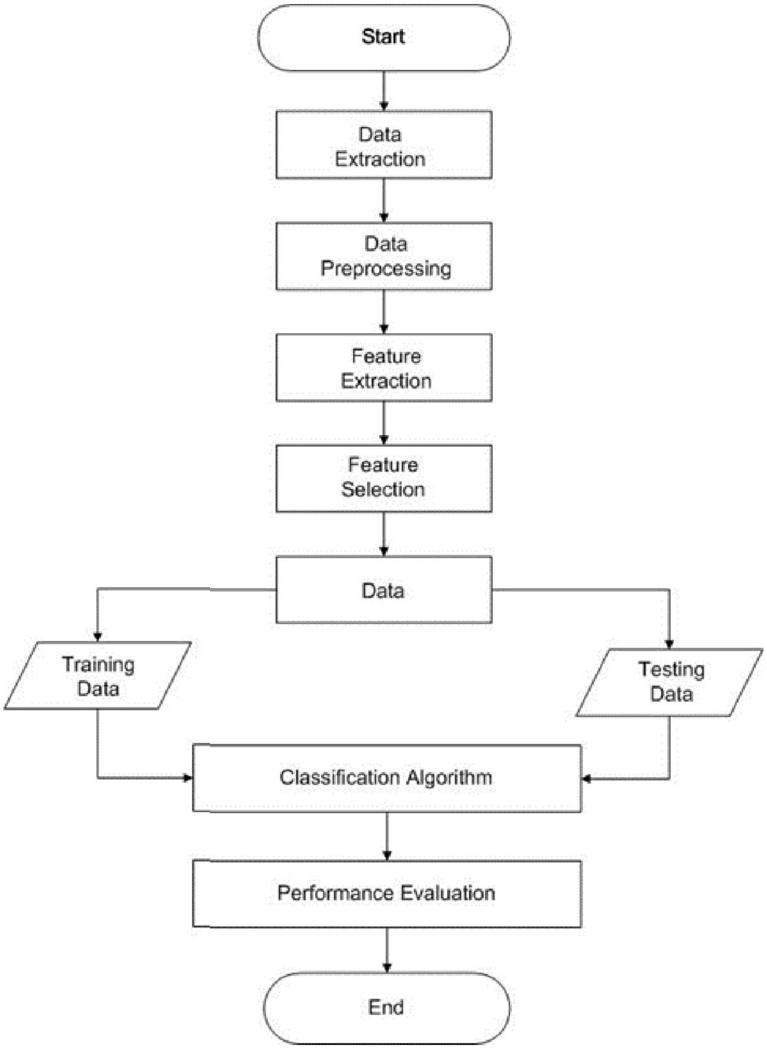

Figure 1: Proposed System Steps

Feature Extraction: Linguistic Feature, semantic feature, psycholinguistic feature, social network feature are extracted to predict personality from tweets posted by user. Feature extraction is based on which kind of prediction is required.

Feature Selection: After extracting features, feature selection algorithm is applied to all the features to find important features. It is not necessary that all the extracted features help in predicting personality. Feature selection select feature which can affect more in prediction.

Classification: Data is split into train and test data set and feed to classifier. Classification algorithm identifies class for new unknown dataset which is verified with the help of test data. Decision tree, Bayesian Classifier, Neural network, K-nearest neighbor, Support Vector Machine. We need to classify data in five personality traits (Openness, Extraversion, Agreeableness, Conscientiousness and Neuroticism)

IV. Result Analysis

Here we have compared accuracy of random forest and KNN algorithm with BOW (bag of words), TF-IDF and State transition feature selection algorithm.

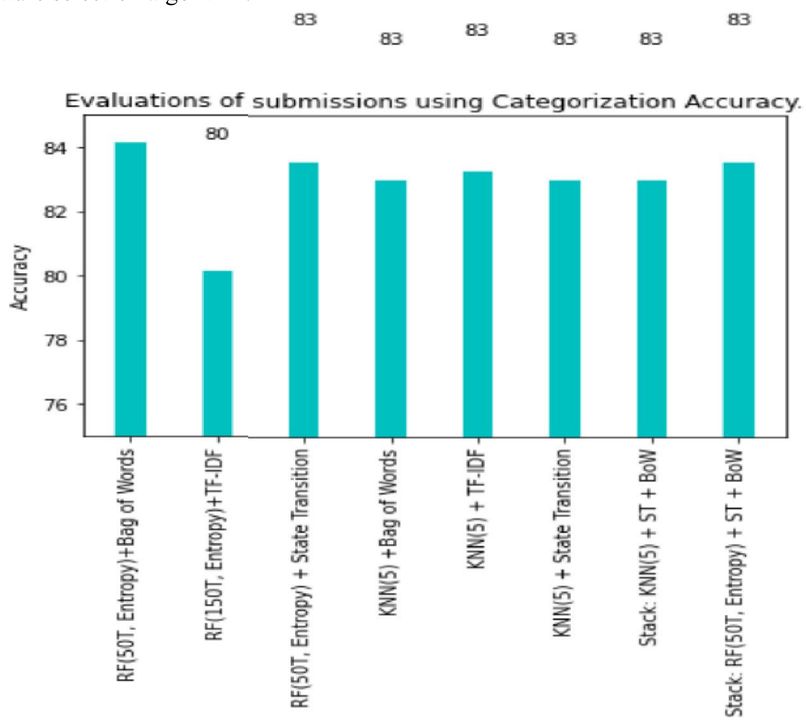

Figure 2: Result Analysis: Accuracy with various algorithm is represented on graph

V. Conclusion

In this study, it is concluded that though a large dataset is being generated by users on social media predicting personality is a complex task. We have also reviewed techniques used by various authors to predict personality traits. According to this paper KNN using TF-IDF is the best model to predict personality. Predicting personality can be useful to plan marketing strategy, study mental health, recruit applicants, education counselling. User's personality is related to his/her social media behavior. Personality prediction from social media is very easy, cost effective and accurate. It requires less time to give results.

For future work we can check double meaning statements, like positive words are written but the meaning is negative. Also, sometimes users can copy random posts or quotes from the internet which do not play any role in predicting personality. We need to ignore it.

References

[1] S. Arjaria, A. Shrivastav, A. S. Rathore and Vipin Tiwari "Personality Trait Identification for Written Texts Using MLNB." R. K. Shukla et al. (eds.), Data, Engineering and Application Springer (2019).

[2] NadeemAhmad, et al. "Personality Assessment using Twitter Tweets"Procedia Computer Science 112 (2017) 1964–1973

[3] P.S.Dandannavar, S.R.Mangalwede, P.M.Kulkarni. "Social Media Text - A Source for Personality Prediction." IEEE, 2018.

[4] Helly. N. Desai, Prof. Rakesh Patel "A Study of Data Mining Methods for Prediction of Personality Traits." International Conference on Smart Electronics and Communication (ICOSEC 2020)

[5] Katiyar Himdweep Walia Sanjay Kumar. "Personality Classification System using Data Mining." International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO 2020)

[6] Joel Philip, Dhvani Shah, Shashank Nayak, Saumik Patel and Yagnesh Devashrayee "Machine Learning for Personality Analysis Based on Big Five Model." Data Management, Analytics and Innovation,Advances in Intelligent Systems and Computing 839 Springer (2019).

[7] Willy, Erwin B. Setiawan, Fida N. Nugraha "Implementation of Decision Tree C4.5 for Big Five Personality Predictions with TF-RF and TF-CHI2 on Social Media Twitter" International Conference on Computer, Control, Informatics and its Applications(2019)

[8] Aditi V.Kunte, Suja Panicker "Using textual data for Personality Prediction:A"Machine Learning Approach" 4th International Conference on Information Systems and Computer Networks (ISCON) Nov 21-22 2019.

[9] Ahmed Al Marouf , Md. Kamrul Hasan, and Hasan Mahmud "Comparative Analysis of Feature Selection Algorithms for Computational Personality Prediction From Social Media" IEEE Transactions On Computational Social Systems (2020).

[10] Srilakshmi Bharadwaj, Srinidhi Sridhar, Rahul Choudhary, Ramamoorthy Srinath "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach" IEEE Xplore(2020).

[11] Lim, A (2020, June 15). The big five personality traits. Simply Psychology. https://www.simplypsychology.org/big-five-personality.html

[12] Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), p.12.https://www.kaggle.com/kazanova/sentiment140

[13] Ruthu S Sanketh,(2020, Dec 3) Text Preprocessing with NLTKhttps://towardsdatascience.com/text-preprocessing-with-nltk-9de5de891658

DOI: 10.48175/568