Real Or Virtual: Personality Traits Analysis For Young Adults On The Basis Of Tweets On Twitter Using Big Five Model

Dr. Kusum Lata Jain[1], Dr. Shivani Gupta[2], Dr. Suyesha Singh[3], Smarnika Mohapatra[4], Archit Shrama[5], Sukriti Sharma[6]

[1,3,5,6] Maniapl University Jaipur, Jaipur,

[2] Vellor Institute of Technology, Chennai

[1] kusumlata.jain@jaipur.maniapl.edu, [2] shivani.gupta@vit.ac.in, [3] suyesha.singh@jaipur.manipal.edu,

[4]smaranika.mohapatra@poornima.edu.in, [5] architsharmadli@gmail.com, [6] sukriti.arora.2017@gmail.com

ABSTRACT

The open-minded use of social platforms provides an area of study where the user thoughts on social media can be used for analysis of their personality. In this paper analysis of personality traits of twitter users on the basis of text tweets is conducted. An age group of young adults (18-25) is selected for the study as this age group is mostly active on the social platform. Big five model of personality traits is used to analysis of personality. Rule-Based Natural language processing is used to analysis on python. A subset of twitter user also tested with Big Five psychological test to test their personality traits in real world and comparison for the same is also performed.

Keywords

BIG FIVE, Personality Traits, twitter, Rule Based Natural Language

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

Social media is a new medium communication through which users connect each other on online communities to interact, share personal and professional information, ideas, messages and other contents. It allows people with same interest to collaborate this, makes it even more powerful. With the rise of social media, distribution of information has been emulated. Users have different ways to interact and tend to exhibit different behaviours such as sharing, commenting, liking, posting and befriending very conveniently. By observing behaviours through social media one can analyse individual's psychological behaviour. Data from online networking sites provides us the ability to capture user data and to know a person's character has grown with the increased use of social media like Facebook, Twitter and Google etc. This project focuses on analysing personality traits based on their tweets on popular microblogging platform Twitter. The project looks into user's interests and track user's activities based upon which we categorize user into particular category of behaviour. The science of behaviour analysis has made discoveries proven very useful in dealing with social behavior such as drug taking, healthy eating, workplace safety, education, and the treatment of pervasive developmental disabilities.

Twitter is specifically chosen because the tweets/content of an individual can easily be accessed and viewed if the owner of the account has made his or her account public. Another important criteria for this study was to collect tweets of individual who were above the age of 18 -25 considered as young adults. A specific feature set is used and the algorithm explores the correlations between each of the feature sets and personality traits. The Big 5 or OCEAN model illustrates 5 personality traits such as Openness, conscientious, extraversion, agreeableness and neuroticism. A cautionary and an important note to be taken here is that the selected individual has to be twitter user who tweets regularly. Finally, features are examined with the highest correlations and apply machine learning algorithm to explore the degree to which one can predict personality traits from tweets from Twitter. The paper is organized as follows section 2 BIG FIVE personality traits analysis models Section 3 Behavior Analysis, Section 4 results section 5 conclusion.

Big Five Model

Big 5 personality model is used widely to predict the behaviour of an individual with 5 basic dimensions of personality. Christopher J Soto[1] proposed Big 5 personality traits, which is prejudiced by both biological and environmental factors.

Individual differences in personality change are somewhat heritable suggesting a biological influence, but have also been linked with a variety of life experiences which suggesting an environmental influence. Literature also explains that Evidence for the Big Five model comes from research examining both everyday language use and formal personality tests. Psycho lexical studies shows the comparison in personality-descriptive language across cultures—have found that most of the world's languages include words synonymous with each of the Big Five, and that the Big Five structure can be recovered from personality ratings made using representative sets of personality-descriptive adjectives in these languages. Psychologists believe in five basic dimensions of personality, referred to "Big 5" personality traits[11]. Following figure shows the personality traits dimensions for BIG five Model

- Extraversion: energetic/reserved
- Agreeableness: friendly/Selfish
- Openness: curious/cautious
- Conscientiousness: organized/careless

- Neuroticism: sensitive/Very relaxed

Figure 1: Big 5 Personality Model

Bogdan Batrinca and Philip C. Treleaven in [2] observes that Computation social science uses quantitative techniques to investigate question which can be used for Social media. In the paper, as the number of user and frequency of use increased on the social media it is become a very large data set and as user use it to share their personal life and event and stories which shows human behavior making new prospects to understand individuals, groups and society. As stated by author in [3], every individual possess different personality trait and user's real and virtual behavior are different, with rise of social media use. Information sharing has been matched observed that individual behavior can be considered as:

- User-user behavior
- User- entity behavior
- User- community behavior

There are many different, new and easy automatic ways to collect this large data from the social media platform for analysis and create new conclusions. In [4] proposed many ways to perform behaviour analysis.

Computational Linguistics includes various techniques can be applied for this analysis like supervised learning, unsupervised learning and Naïve Bayes algorithm. Literature suggest that behaviour analysis is a trending research as it allows people to understand themselves and others very well. In [5] proposed used rule-based Natural Language Processing to improve disease normalization in biomedical text and explained how Rule-based NLP can be used to extract useful information from the unstructured text. Mukul Aggarwal and Ashwini Kumar[6] proposed sentiment analysis and feature extraction using Rule Based Model (RBM) and explains that comprehensive sentiment lexicon, created by user data, is difficult process and may introduce possibilities of having error. Therefore, usually researchers depend on existing lexicons as primary resources. RBM has advantage of easy computation and more accurate than other approaches. No training data set is required so it can be used for different domain as well as for individual data set. The lexicon and its rules can be easily accessed by the user. It is not hidden within a machine access. It is easily inspected, understood, extended, or modified. This make it very rule based system to use with new models and domains[8,9].

Methodology &Framework

Behaviour analysis is analysis of behavior of individuals. The study is conducted for analysis of personally traits using tweets on Tweeter. Python is used as a programming language. Visual Studio Code is used as an environment to run. Following methodology is used for the same:

A. Data Collection

Form a data set of 1500 tweets using 25 tweets of 60 individual person. These tweets are picked up as raw aka not refined or touched to maintain its authenticity. These tweets are stored in an excel file. Twitter API's were learnt but couldn't be used to move the data due to the age restriction defined in twitter legal documentation. Age is a criteria through which tweets cannot be extracted. Tweets can be extracted through location, name, key words, trends etc but not through age.

B. Pre-Processing of Data

Link and clean the data set for pre-processing and processing. Python is used as a programming language to implement NLP (Natural Language Processing) functions onto the data set to refine the data to purest form to carry out the desired operation. The given data set's vales are refined in the following way

- Every word converted into lower case.
- Removal of punctuations and exclamations.
- Removal of stop words
- Lemmatization: Lemmatization refers to the process of getting the root word and removing the tense of word. For example studying becomes study. If this function is not implemented the words; studying and study which carry the same meaning are treated as different.
- Removal of numbers and emoticons
- Removal of words like dtype, object and NaN

•Tokenization: It's easier to work on list hence the conversion from data frame to list. Since the individual list has only tweets relating to the single person, it becomes easier for us to tokenize and compare. word tokenize: is an in-built method which converts words into tokens from nltk package.[7]

- Data frame to list: Techniques used to compare tokens with keywords based of personality prediction Correction of words.

C. Keyword matching

Creating an algorithm using Python and machine learning fundamentals to classify the posts into a personality model, The Big 5. A basic approach which is being followed is Rule based NLP. The rule-based system is defined as by using available knowledge or rules. A bag of words containing the dictionary of words that fall into the category of OCEAN was created. A bag of words is a simplified version of NLP and information retrieval[9]. The Big 5 model is applied on the dataset and the posts are further categorized into OCEAN with a count of each personality trait against every individual. We created bag of words of individual Big 5 trait – Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. These will act as the ground truth for the Rule

Based Classifier. Examples are provided in figures 2

Figure 2: Rule based classifier for ageeable

These bag of words are converted into lists for the classifier as

```
agreeableness = open("agreeableness.txt").read().split()
conscientious = open("conscientious.txt").read().split()
extraversion = open("extraversion.txt").read().split()
openness = open("openness.txt").read().split()
neuroticism = open("neuroticism.txt").read().split()
```

Figure 3: creation ofbag of words for Big FiveModel

Matching of the the tokens of an individual tweets with the keywords of an individual traits from Big 5 model is performed after classification. Conclusion from the individual's person personality on social media with the help of keywords.These conclusion will be shows as %.

D. Psychological test

A psychological test is conducted for selected participant in person and derive his personality traits from the inputs.

E. Comparison of Analysis

Compare the result of the personality traits derived through tweets versus the personality traits derived through the Big 5 model form.

Behavior Analysis &Result

Analysis shows the user behaves very differently on online social media platform from real word. Psychological test is performed for 25 people. As microblogging site twitter is mostly used by celebrities. After approaching to the people all are not comfortable with psychological test. So less number of test conducted. Following figure shows the results of the two participants for psychological test. Participant 1 and 2

Behavioral analysis shows

E = 20 + 4- 2+ 5 - 2 + 2 - 2 + 4 - 2 + 4 - 3 = 28

A = 14 - 5 + 4- 1 + 5 - 1 + 4 - 1 + 4 + 5 + 4 = 32

C = 14 + 4 - 2 + 4 - 2 + 3 - 2 + 4- 2 + 3 + 2= 26

N = 38 - 2 + 2 - 3 + 5 - 2- 1 - 3 - 1 - 1- 1=32

O = 8 + 3 - 2 + 4 - 2 + 3 - 1 + 4 + 3 + 3+4 = 27

Figure 4: Participants1 psychological Test

$E = 20 + {}_{(1)}\underline{2} - {}_{(6)}\underline{4} + {}_{(11)}\underline{2} - {}_{(16)}\underline{2} + {}_{(21)}\underline{3} - {}_{(26)}\underline{3} + {}_{(31)}\underline{3} - {}_{(36)}\underline{5} + {}_{(41)}\underline{5} - {}_{(46)}\underline{5} = 21$

$A = 14 - {}_{(2)}\underline{2} + {}_{(7)}\underline{3} - {}_{(12)}\underline{3} + {}_{(17)}\underline{4} - {}_{(22)}\underline{1} + {}_{(27)}\underline{5} - {}_{(32)}\underline{2} + {}_{(37)}\underline{2} + {}_{(42)}\underline{4} + {}_{(47)}\underline{3} = 27$

$C = 14 + {}_{(3)}\underline{3} - {}_{(8)}\underline{1} + {}_{(13)}\underline{5} - {}_{(18)}\underline{3} + {}_{(23)}\underline{3} - {}_{(28)}\underline{1} + {}_{(33)}\underline{5} - {}_{(38)}\underline{2} + {}_{(43)}\underline{4} + {}_{(48)}\underline{3} = 30$

$N = 38 - {}_{(4)}\underline{2} + {}_{(9)}\underline{4} - {}_{(14)}\underline{5} + {}_{(19)}\underline{3} - {}_{(24)}\underline{2} - {}_{(29)}\underline{2} - {}_{(34)}\underline{1} - {}_{(39)}\underline{2} - {}_{(44)}\underline{3} - {}_{(49)}\underline{1} = 27$

$O = 8 + {}_{(5)}\underline{3} - {}_{(10)}\underline{1} + {}_{(15)}\underline{5} - {}_{(20)}\underline{1} + {}_{(25)}\underline{4} - {}_{(30)}\underline{1} + {}_{(35)}\underline{2} + {}_{(40)}\underline{2} + {}_{(45)}\underline{5} + {}_{(50)}\underline{5} = 31$

Figure 5: Participants2 psychological Test

Following table 1 and 2 shows comparison for personality traits on online and real world behavior for user1 and 2.

Table 1: comparison for personality traits on online and real world behavior for participant 1

Personality trait	The results of the selected participant as per his behaviour in the outside world:	The results of the selected participant as per his behaviour on social media; Twitter:
Extraversion:	19.3%	0%
Agreeableness:	22.06%	0%
Consciousness:	17.95%	40%
Openness:	22.06%	20%
Neuroticism:	18.62%	40%

Table 2: comparison for personality traits on online and real world behavior for Participant 2

Personality trait	The results of the selected participant as per his behaviour in the outside world:	The results of the selected participant as per his behaviour on social media; Twitter:
Extraversion:	19.3%	0%
Agreeableness:	22.06%	0%
Consciousness:	17.95%	40%
Openness:	22.06%	20%
Neuroticism:	18.62%	40%

The comparison show s that participants 1 appears to display more consciousness and neuroticism on Twitter as compared to the behaviour in the outside world.
The participant's feed of 25 tweets lacks words which could be related to extraversion and agreeableness. The participant scores roughly the same in the area of Openness. Following figure shows the comparison for participants 1.

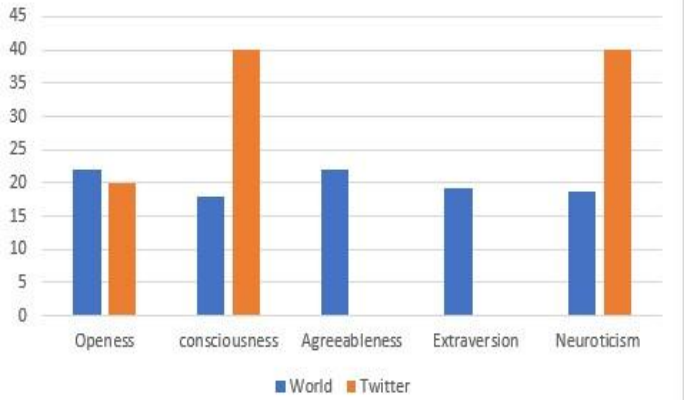

Figure: Bar graph of the participant 1

Following figure shows the results for participant 2 shows more extraversion behaviour on Twitter as compared to the behaviour in the outside world. The participant displays equal amount in the consciousness and Openness. The participant's feed of 25 tweets lacks words which could be related to agreeableness. The participant roughly scores the same in the area of Neuroticism. . Following figure shows the comparison for participants 1.

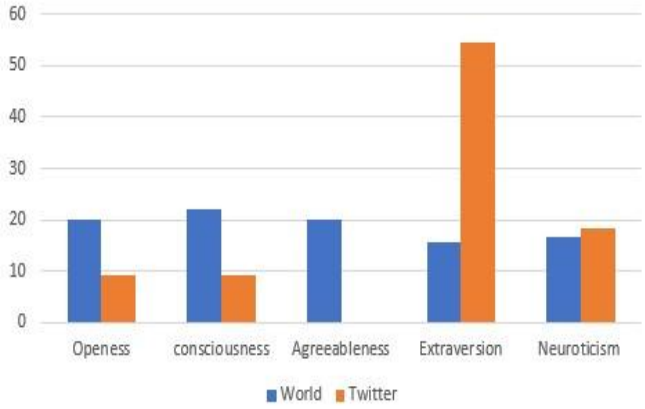

Figure 26: Bar graph of second participant

Analysis shows that users shows very differently on virtual world then real world. And trends shows that young adults are more conscious on online then real world. 69 percent user behave differently in real and online world. All participant behave differently in one or other personality.
The behaviour displayed on socials is completely different from the behaviour displayed in the outside world. 80% of the time, traits would likely not match against each other. No definite conclusions can be drawn by comparing both of these results against each other however both of these results can be collectively used for a singular research of a person. One trait is likely to match between the two results. The given experiment's results can be strengthened further by increasing the database of an individual.

References

[1] Soto, C. J. (2018). "Big Five personality traits" In M. H. Bornstein, M. E. Arterberry, K. L. Fingerman, & J. E. Lansford (Eds.), "The SAGE encyclopedia of lifespan human development" (pp. 240-241), 2018.

[2] Batrinca, Bogdan, and Philip C. Treleaven. "Social media analytics: a survey of techniques, tools and platforms." Ai & Society 30.1 (2015): 89-116.

[3] S.D. Kularathne, R.B. Dissanayake, N.D. Samarasinghe, L.P.G. Premalal and S.C. Premaratne, "Customer Behavior Analysis for Social Media" [Vol-3, Issue-1, Jan-2017] ISSN : 2454-1311

[4] Michael M. Tadesse, Hongfei Lin, Bo Xu and Liang Yang, "Personality Predictions Based on User Behavior on the Facebook

[5] Social Media Platform" [Date of publication – 17th Oct 2018] DOI: 10.1109/ACCESS.2018.2876502

[6] Hajra Waheed, Maria Anjum, Mariam Rehman and Amina

[7] Khawaja, "Investigation of user behavior on social networking sites" Date of publication – 2nd Feb 2017] DOI: 10.1371/journal.pone.0169693

[8] Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E. M., & Kors, J. A. (2014). Knowledge-based extraction of adverse drug events from biomedical text.BMC bioinformatics, 15(1), 1-8. [6] Sentiment Analysis and Feature Extraction Using Rule-Based Model (RBM) Raghavendra Kumar Dwivedi, Mukul Aggarwal, Surendra Kr. Keshari and Ashwini Kumar-Springer Nature Singapore Pte Ltd. 2019 S. Bhattacharyya et al. (eds.), International Conference on Innovative Computing and Communications, Lecture Notes in Networks and Systems 56. www.analyticsvidhya.com/blog/2018/02/the-different-methods-de al-text-data-predictive-python/#

[9] Natural Language Processing Alexander Gelbukh National Polytechnic Institute (IPN), Mexico 2005 IEEE.

[10] Rule-base Information Extraction is dead! Long Live Rule-based Information Extraction Systems Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 827–832, Seattle, Washington, USA, 18-21 October 2013. c 2013 Association for Computational Linguistics. www.researchgate.net/publication/319119259_Human_Behavioral_Analysis_Using_Evolutionary_Algorithms_and_Deep_Learning www.verywellmind.com/the-big-five-personality-dimensions-27 95422