International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Five Factor Model of Personality Trait Analysis on Twitter Data using Benchmark Classifier

Sayeda Umera Almas, Puttegowda D

Abstract*: Health and wealth of the society is directly proportional to the activities conducted based on healthy personality traits of the citizens and their social behavior's. Hence, the diagnosis and its preventive measures play a very important role and may be challenge for medical and engineering domains. The proposed paper is trying to analyze the personality traits based on Five Factor Model by processing the twitter dataset. The classification models are trying to give number of solutions corresponding to large amount of data (Big data). Classification technique may predict the personality qualities of the user based on their interaction with the system. This diagnosis may support the society in bringing up healthy environment for better lifestyle of everybody.

***Keywords* : Personality traits, Classification models, Twitter data analysis, Natural Language Processing.**

I. INTRODUCTION

Five Factor personality traits Model is a psychological solution emphasizing mainly on the personal qualities and differences that exists in social behavior [1]. These day's social media is generating enormous amount of data through the user interactions. The social media applications have provided number of ways for the user to share their views and express their emotions through images, text, voice and so on[2]. These could be the sources to analyze various perspectives of the users concerned to their emotions. These emotional aspects certainly connected to the behavior of the users. The interactions with the applications through text may yield the certain features. These features may classify the personality of the user among predefined five factor personality traits. Since the dataset is not available for the model, unsupervised techniques are employed for clustering and classification of the samples.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Sayeda Umera Almas*, Research Scholar, ATME College of Engineering, Mysuru, Visvesvaraya Technological University, Belagavi, India, email:umera.almas@gmail.com
Dr. Puttegowda D, Professor and Head, Department of Computer Science & Engineering, ATME College of Engineering, Mysuru, Visvesvaraya Technological University, Belagavi, India, email:pgdatme@gmail.com

II. LITERATURE SURVEY

A. Five Factor personality traits Model

Psychologists have introduced five basic personalities that help to categorise human personality in a simple way.

It is also known as Big Five Personality Traits Model [3], It is also called as OCEAN model. It is based on Five Factor Model of personality that describes the qualities of the person, which helps to identity his or her behavioral properties. The author has introduced Five Factor Model [4] in the year 1990, which gained a wide acceptance across the globe. The basis of personality traits, the Five Factor Model is found to be practical and its applicability is well supportive to cross observers and different culture. There are five major personality traits that influence the way we behave, the career we choose and the lifestyle we live. These are the inherent characteristics dominate human existence. Each person displays a characteristic personality trait that prevails around his or her life. Also quite a few will display these personality traits variant degrees. One personality trait dominates in each one of us that reflects our character [5].

1. *Neuroticism(N):* People who show high degree in neuroticism are least stable emotionally and they tend to react to little things and get upset easily. Also studies have shown that these people will be suffer from depression. Keywords: upset, stress, negative mood, angry, anxious, depression, moody, irritate [6].

2. *Extraversion(E):* People with high degree of extraversion are more tend to social behavior's. They enjoy participating in social activities, interact with people, travel and they are extremely friendly. They feel uncomfortable spending time alone. Keywords: social, strong, friendly, party, talkative, discussion, lovely

3. *Agreeableness(A):* Politeness and compassion are the hallmarks of the people exhibit this personality trait. They enjoy helping others and it is associated with good behavior and they are good hearted people, extremely cooperative and trustworthy. Keywords: politeness, compassion, help others, good, trustworthy, positive, interactive, empathy, communication, respect, unselfishness

4. *Conscientiousness(C):* These people are goal-oriented with highest levels of thoughtfulness. They will have exceptionally organization skills and plan everything well in advance. They are disciplined, punctual, deliberate and careful with respect to further activities to take up. Keywords: soft discipline, plan, organize, Tasks, responsible, achieve goals, descent, self-control, work hard, perfection, workholic, reliability, strict, clean

5. *Openness(O):* They are open minded and free to accept any challenges. They are adventures and move out of their comfort zone and experience new things. They sit and work hard to enhance their knowledge. They enjoy finding a solution to the creative puzzle. Keywords: Open minded, new idea, adventures, experience new things, learn, creative, innovative, conservative.

B. Research using Twitter data

Twitter is a microblogging service provider by providing a platform for live communication to its users. Author [7] has utilized twitter messages for the sentiment analysis. Author [8] has referred twitter data for the sentiment analysis based on the top 30 major events.

C. Clustering and Classification

Now-a-days the clustering and classification models have been solved number of complex problems in the society. These are mainly used in Machine Learning

III. METHODOLOGY

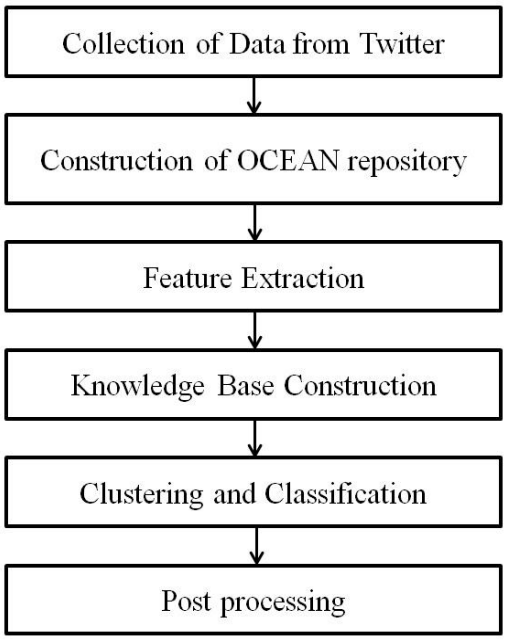

Fig. 1.Methodology of the proposed system

Figure 1 shows the methodology of the proposed paper. This methodology mainly emphasizes on the extraction of desired behavioural keywords from the twitter data and constructs knowledge base. Knowledge Base is used to classify the test sample with respect to Five Factor Personality Traits model.

A. Collection of Data from Twitter

Tweepy is a python class or API designed to access Tweets from Twitter application [9]. Tweepy supports for mining the required data.

B. Construction of OCEAN repository

This step utilizes Multi-Construct IPIP Inventories [10] for the keyword collection of the major personality traits from OCEAN model. This OCEAN repository is used for the creation of hidden layer in doc2vec[11] algorithm as shown in Fig. 2. OCEAN repository will be consisting of five sub repositories as per the OCEAN five factor definitions.

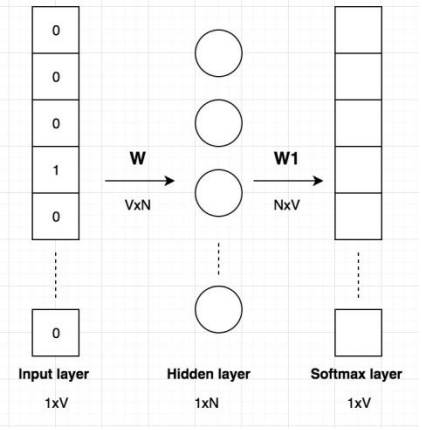

Fig. 2.: The process of doc2vec

C. Feature Extraction and Selection

Features define the sample and its characteristics or behaviour. Following are the certain features considered in this paper.

i. doc2vec [12]: This is one of the existing feature from the literature and it is based on a popular word embedding's algorithm i.e., word2vec. Word2vec and doc2vec are the context based word embedding algorithms. Word2vec algorithm can convert a word to vector whereas doc2vec converts a document to vector. These vector are used for the estimation of the similarity between two different text components. The probability of the similarity of the given document corresponding to the OCEAN repository is estimated using Softmax algorithm.

Word2vec has two architectures namely CBOW (Continuous Bag of Words) and Skip-gram.

CBOW is the method of predicting target word based on number of words. Skip-gram is the process of predicting multiple of words based on a given single word as shown in Fig 3.

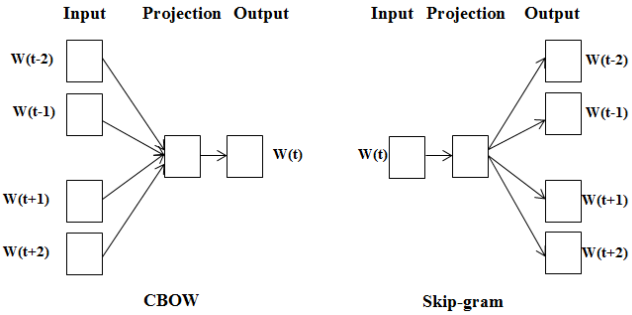

Fig. 3.: Architecture of CBOW and Skip-gram

Softmax evaluation results the probability of the context semantics corresponding to the keyword.

i. OCEAN-Coefficient: this is the proposed feature, which will focus mainly on the presence of the keywords and their thesaurus as per the OCEAN descriptions.

Extraction of the behavioural keywords from the tweets is rendered using the following algorithm.

Algorithm: Extraction of behavioural keywords

Step 1: Removal of repeated or unimportant words from the tweets using TF-IDF

Step 2: Assignment of Weightage as per OCEAN repository

Step 3: Estimation of the OCEAN coefficient

D. Knowledge base construction using Clustering

Two features namely OCEAN and doc2vec coefficients are estimated for about hundreds of an individual tweets. Since, the samples are not labelled, the unsupervised technique K-means[13] is applied on these samples by considering k=5. The process of clustering and labelling these samples as either O, C, E, A or N. This labelled data can be regarded as Knowledge Base.

E. Classification and post processing

Testing process is conducted using k-NN algorithm. k-NN is a supervised classification technique, which labels the test sample based on the knowledge base.

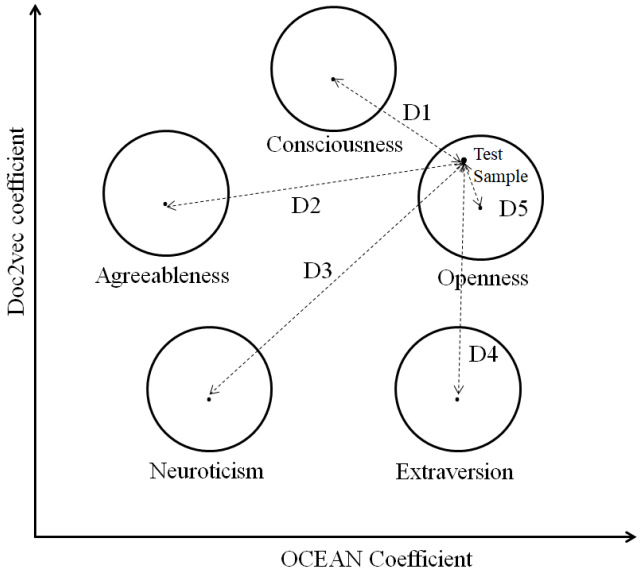

Fig. 4.: Clustering and Classification of Test sample based on OCEAN and Doc2vec Coefficients

Fig 4 depicts the distances D1, D2, D3, D4 and D5 from centroids of the five clusters to the test sample. The degree of the similarity can be easily estimated using these distances. Distances are estimated using (1)

$$D1 = SQRT((x1-x2)^2 + (y1-y2)^2) \quad (1)$$

The degree of the similarity corresponding to each behaviour factor is estimated using the (2)

$$Degree\_sim = D_i / \text{Sum of all the distances} \quad (2)$$

Where i is from 1 to 5

IV. RESULTS AND DISCUSSIONS

Table I shows the example data from Multi-Construct IPIP Inventories [14]. This data is used for the extraction of doc2vec feature between Table I data with Tweets collected

Table I: Example data from Multi-construct IPIP inventory

Openness	**Enjoy hearing new ideas, Have a rich vocabulary….**
Consciousness	Carry out my plans, Pay attention to details…
Extraversion	Feel comfortable around people, Make friends easily…..
Agreeableness	Respect others, Accept people as they are…
Neuroticism	Worry about things, Get stressed out easily….

Table II depicts the around hundred samples along with features extracted and labelled, hence called as Knowledge Base for the further classification.

Table II: Knowledge base for behavioural analysis

Tweet	**OCEAN coefficient**	**Doc2vec coefficient**	**Label**
T1	0.54	0.33	3
T2	0.58	0.95	1
.....			
T100	0.96	0.23	2

Figure 5 shows the graphical representation of the samples, their clusters and centroids. Also the graph shows the test sample indicated with + mark.

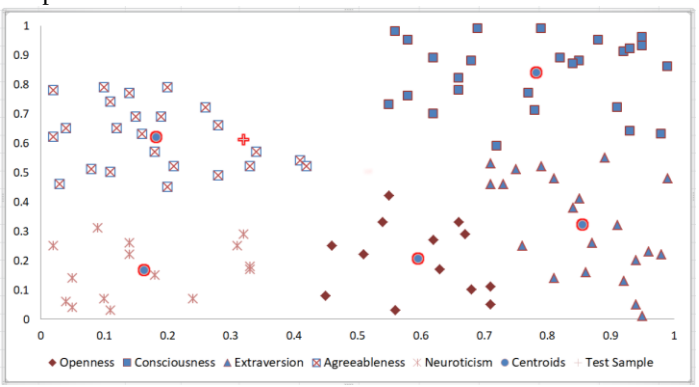

Fig. 5.: Clusters using Kmeans algorithm

Table III indicate the centroids of the five clusters. Table IV shows the two test samples and their distances from the centroids of all the five clusters.

Table III: Centroids of the five clusters

	x1	**y1**
Centroid0	0.60	0.20
Centroid1	0.78	0.84
Centroid2	0.86	0.32
Centroid3	0.18	0.62
Centroid4	0.16	0.17

Table V demonstrates the distribution of the distances shown in Table IV. This distribution is helpful to indicate the degree of the participation of the test sample in each cluster.

Table IV: Test samples and their distances from centroids of five clusters

	OCEAN	doc2vec	D1	D2	D3	D4	D5
Test sample1	0.32	0.61	0.49	0.52	0.61	0.14	0.47
Test sample2	0.59	0.48	0.28	0.41	0.31	0.43	0.53

Table V: Distance distribution of two test samples

	OCEAN	doc2vec	D1	D2	D3	D4	D5
Test sample1	0.32	0.61	0.22	0.23	0.27	0.06	0.21
Test sample2	0.59	0.48	0.14	0.21	0.16	0.22	0.27

Figure 6 shows the graphical representation of Table V, it is clearly says that Test sample 1 is more towards the fourth property of OCEAN that is Agreeableness but far from Extraversion. Similarly, Test sample 2 is more towards Openness but away from Neuroticism.

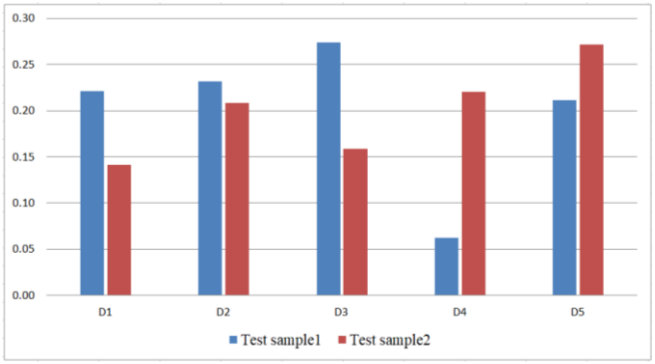

Fig. 6.: Degree of personality traits of two test samples

V. CONCLUSION

The involvement of technology in personality trait analysis certainly motivates the researcher to think about other resources of input for the research. In this context, the proposed paper has tried to analyze the twitter data of an individual for the analysis of their personality traits. Also the paper has employed machine learning or classification models for the same. These attempts are very much essential in the current information technology generation for the number of inferences and declarations. These may lead to understand ourself for the betterment of the society.

ACKNOWLEDGMENT

Authors would like to thank the Management, Principal and all the teaching, Non-Teaching faculty members of *ATME College of Engineering, Mysuru, India* for supporting and encouraging our research work.

And we gratefully thank the Visvesvaraya Technological University, Jnana Sangama, Belagavi for encouragement and support extended to this research work

REFERENCES

1. Chamorro-Premuzic, T. (2012). Personality and individual differences. West Sussex: BPS Blackwell.
2. Z. Wang, C. S. Chong, L. Lan, Y. Yang, S. Beng Ho and J. C. Tong, "Fine-grained sentiment analysis of social media with emotion sensing," 2016 Future Technologies Conference (FTC), San Francisco, CA, 2016, pp. 1361-1364.doi: 10.1109/FTC.2016.7821783
3. Goldberg, L. R. (1993). The structure of phenotypic personality traits. American Psychologist, 48, 26–34.
4. Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual Review of Psychology, 41, 417–440.
5. https://www.cleverism.com/big-five-personality-traits-model-ocean-model/, Accessed on 10/01/2020
6. Widiger, T. A. (2009). Neuroticism. In M. R. Leary & R. H. Hoyle (Eds.), Handbook of individual differences in social behavior (p. 129–146). The Guilford Press.
7. Bruns, Axel, et al. "A topology of Twitter research: disciplines, methods, and ethics." Aslib Journal of Information Management (2014).
8. Thelwall, Mike, Kevan Buckley, and Georgios Paltoglou. "Sentiment in Twitter events." Journal of the American Society for Information Science and Technology 62.2 (2011): 406-418.
9. Roesslein, Joshua. "Tweepy." Python programming language module (2015).
10. https://ipip.ori.org/newMultipleconstructs.htm, accessed on 03.01.2020
11. Z. Wang, C. S. Chong, L. Lan, Y. Yang, S. Beng Ho and J. C. Tong, "Fine-grained sentiment analysis of social media with emotion sensing," 2016 Future Technologies Conference (FTC), San Francisco, CA, 2016, pp. 1361-1364.doi: 10.1109/FTC.2016.7821783
12. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
13. Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.
14. Collaboratory, I. P. I. P. "Multi-Construct IPIP Inventories." (2016).

AUTHORS PROFILE

Sayeda Umera Almas an full time Research Scholar under Dr. Puttegowda D, from Computer Science & Engineering Department of ATME College of Engineering, Mysuru, which is affiliated to Visvesvaraya Technological University, Belagavi, India. Earlier she was working as an Associate System Engineer at IBM Bangalore, India. Currently she is working on Natural Language Processing, Machine Learning using Python programming language and she has published several papers in National and International publications.

Dr. Puttegowda D, Professor and Head at Computer Science & Engineering Department of ATME College of Engineering, Mysuru, which is affiliated to Visvesvaraya Technological University, Belagavi, India. He received Doctoral degree in field of Image Processing. He has published many papers and journals in Image Processing, Data mining, Big-Data Analytics, Machine Learning and other area. He has sixteen years of teaching and two years of industry experience. His Research interests are Digital Image Processing, Data mining, Big-Data Analytics, Machine Learning and Pattern Recognition.