Ocean Coefficient: A Feature Extraction Technique for Five Factor Model based Classifications

Sayeda Umera Almas, Puttegowda D

***Abstract*: Natural Language Processing has opened up several avenues in the field of research and developments. It has supported wide variety of applications, but still the opportunities are enormous for the researchers to look into several other aspects in the discovery of new dimensions. In this regard the current paper is trying to introduce a revolutionary feature extraction technique, particularly for the studies/research corresponding to five factor model based behaviour analysis.**

***Keywords* : Personality traits, Feature extraction technique, OCEAN Model, Natural Language Processing**

I. INTRODUCTION

Behaviour is one the major parameter of human beings. This parameter may categorize and label the individual based on the type of personality traits. Researchers from all the domains namely psychology, medical, engineering and others have been proposed and proposing number of personality trait analytical approaches. Image, speech and text [1,2,3,4,5] are the various inputs considered for the behavioural analysis. Five Factor Model (FFM) is one of the globally accepted personality trait analysis model [6]. Openness(O), Conscientiousness(C), Extraversion(E), Agreeableness(A), and Neuroticism(N) are the identified traits under FFM, it is also called as OCEAN Model [7]. The process of clustering and classification mainly involves the selective parameters of the samples. Acceptability of the classification model mainly relies on the applicability of the parameters, which are involved in the process. In this regard, this paper is trying to propose one of the feature extraction technique called as OCEAN coefficient. This can be used as one of the parameter in FFM based classification models.

II. REVIEWOF LITERATURE

A. Five Factor personality traits Model

The literature explores number of psychological inventions in deriving solutions for the analysis of behavioural component an individual. Each of these is defined and derived in their own perspectives to reach the objectives. In this regard, the authors introduced a revolutionary and widely accepted behaviour estimation instrument in the year 1990 called as Five Factor Model. It is also known as Big Five personality Traits Model. FFM consolidates and segregates all the behavioural qualities into five major traits namely Openness(O), Conscientiousness(C), Extraversion(E), Agreeableness(A), and Neuroticism(N) [7].

1. *Neuroticism(N)[8]:* This quality narrates the people who do not analyse the consequences, they speak before think and become angry suddenly on negligible issues. Emotionally they are less stable, react to little things and upset easily. They moody in nature and always will be thinking everything in negative direction.
2. *Extraversion(E*): People belonging to this group will enjoy being in group rather than being alone. They like to be the centre of attention and are socially very active who love to be happy and spent lots of time in social gatherings or social events or parties. They are assertive , energetic ,positive and full of life .
3. *Agreeableness(A):* people belonging to this group are cooperative, warm, considerate ,sympathetic and kind in nature. People with high agreeableness are trustworthy, affectionate, and friendly. They have a great concern for the welfare of others and will be the first to help the needy. They does not care about the offending comments. They have less patience and are annoying.
4. *Conscientiousness(C):* These people are very serious when it comes to work or task. They take others obligations seriously. They are well ordered and does not like easy going. They are well organised, responsible, self-controlled. People with high conscientiousness are self-discipline, hardworking and they put constant efforts to reach their goals. They are perfectionists when it comes to achieving their gaols.
5. *Openness(O):* people belonging to this group loves challenges. They does not restrict themselves and love to try new things and new experiences. People with high openness tend to be open minded, adventurous, innovative.

Revised Manuscript Received on April 21, 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

Retrieval Number: D8395049420/2020©BEIESP
DOI: 10.35940/ijeat.D8395.049420
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication
© Copyright: All rights reserved.

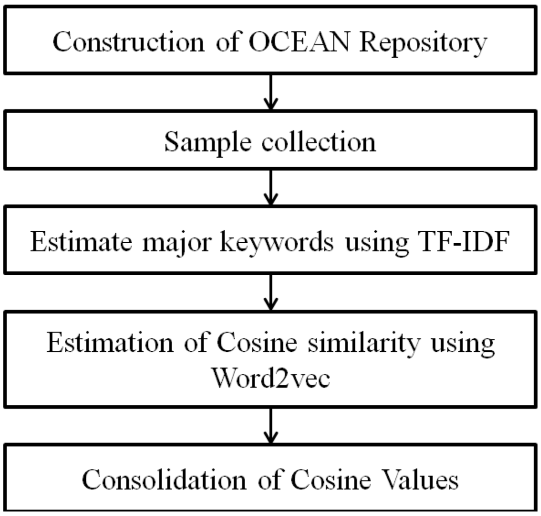
They are good learners and are very creative and conservative in nature. They take others comments positively and try to improve themselves.

B. Natural Language Processing (NLP)

Literature has proved that NLP has wide variety of applications corresponding to the factorization of text samples.

Author [9] has tried to verify the importance of emails based on the subject and content of the email. Here author employs unsupervised technique and found 91% of efficiency using M3 model of word2vec NLP classification. Author [10] has employed AdaBoost classifier to estimate the genre of the text book, so that its right audience can be predicted before to introduce the book into market. Author [11] has tried process the music and its lyrics to classify them into various moods. Also, certain studies [12] have shown the importance of feature selection and their varied efficiencies corresponding to the number of features opted. Papers [13, 14] and many other have been shown the importance of text classification. Author [15] has tried to explore the process of clustering and classification and the importance of feature extraction and their selection.

III. METHODOLOGY


Construction of OCEAN Repository
Sample collection
Estimate major keywords using TF-IDF
Estimation of Cosine similarity using Word2vec
Consolidation of Cosine Values

Methodology of the OCEAN coefficient feature extraction

Figure 1 shows the methodology for the estimation of OCEAN coefficient feature. This involves the identification of major keywords using TF-IDF and finding their cosine values.

A. Construction of OCEAN Repository

Thesaurus [16] is a tool used to identify major keywords orient towards the traits of the OCEAN model. The completeness of the repository can yield precise OCEAN coefficient value. Since, the current paper is trying to propose a model, it is assumed a complete repository by considering certain available keywords.

Table-I: Example OCEAN repository

PROPERTY	KEYWORDS
Openness	New idea, adventures, experience new things, learn, Open minded, creative, innovative, conservative and so on.
Conscientiousness	soft discipline, plan, organize, Tasks, responsible, achieve goals, descent, self-control, work hard, perfection, workholic, reliability, strict, clean and so on
Extraversion	social, strong, friendly, party, talkative, discussion, lovely
Agreeableness	politeness, compassion, help others, good, trustworthy, positive, interactive, empathy, communication, respect, unselfishness and so on..
Neuroticism	upset, stress, negative mood, angry, anxious, depression, moody, irritate and so on..

B. Sample collection

Tweets pertaining to any individual can be considered as a sample. Tweepy[17] python API is employed to extract tweets from the twitter.

C. Extract behavioural keywords from sample

Term Frequency – Inverse Document Frequency (TF-IDF) is a mechanism to identify important keywords from the given input text based on the corpus provided [18]. Here OCEAN repository as shown in Table I can be used as corpus for finding behavioural keywords from the sample.

D. Estimation of cosine similarity using word2vec

Word2vec[19] is mechanism to estimate cosine similarity of the given keyword against the given text. Figure 2 shows two architectures of word2vec in its cosine similarity estimation. This paper follows CBOW (Continuous Bag of Words) architecture.

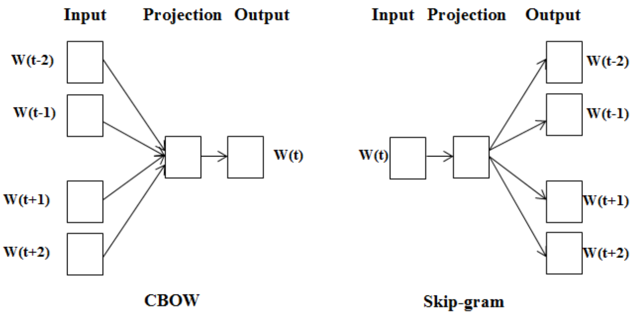


Fig. 1.word2vec architetures

Cosine similarity evaluation results the probability of the context semantics corresponding to the keyword.

E. Consolidation of Cosine Values

Cosine similarity is estimated for each and every keyword of the OCEAN repository against the given sample (tweet of text). Algorithm 1 consolidates the cosine values of each keyword under OCEAN repository.

Algorithm 1: Consolidation of cosine values to estimate OCEAN coefficient

Step1: Initialize oceancoefficient = 0
Step 2: Estimate the maximum cosine value among all keyword cosine values under each property of OCEAN
max_p = maximum (cosine_value(keywords_property))
Step 3: for each property

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

```
If (property==O && max_o>0)
    oceancoefficient = oceancoefficient + 1
If (property==C && max_c>0)
    oceancoefficient = oceancoefficient + 2
If (property==E && max_e>0)
    oceancoefficient = oceancoefficient + 4
If (property==A && max_a>0)
    oceancoefficient = oceancoefficient + 8
If (property==N && max_n>0)
    oceancoefficient = oceancoefficient + 16
```

IV. RESULTS AND DISCUSSIONS

As an example the following sample is considered for the estimation of OCEAN coefficient.

Sample: “we have to plan and organize our parties, celebrations, excursions etc. so that everybody can retain strong relationship”. Table II shows the estimated cosine values of the keywords considered under each property of OCEAN.

Table-II: cosine values of the keywords

Openness		Conscientiousness		Extraversion	
idea	0	plan	0.23	social	0
adventure	0	organize	0.58	strong	0.2
learn	0	Tasks	0	friendly	0.4
……		…….		……..	

Agreeableness		Neuroticism	
help	0	stress	0
good	0	angry	0
polite	0	anxious	0
……..		………	

Table III depicts the consolidation and consideration of the parameters. Since, the max_c and max_e values are greater than 0, their corresponding values as per the algorithm are added to coefficient. Thus the consolidated value can be regarded as OCEAN coefficient. The estimated coefficient is considering all the parameters in its values. The proposed algorithm can consider all the OCEAN parameters whenever the sample exhibits more than one parameter.

Table-III: Consolidation of coefficient

max_parameter	max value	value to be added
max_o	0	0
max_c	0.58	2
max_e	0.4	4
max_a	0	0
max_n	0	0
OCEAN coefficient =		**6**

V. CONCLUSION

Indeed there is no end for the research and developments, innovative thoughts, new dimensions. This paper is one such attempt to define a new feature corresponding to behavioural analysis. Word2vec is used to estimate cosine similarities between the selected keywords with the sample. The proposed algorithm is either verifying 0/1 basis parameter addition. This can be extended to continuous values by adopting new methodologies. The proposed feature can be utilized in the clustering and classification models of the behavioural analysis.

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

REFERENCES

1. Vanderlind, W. M., Millgram, Y., Baskin-Sommers, A. R., Clark, M. S., & Joormann, J. (2020). Understanding positive emotion deficits in depression: From emotion preferences to emotion regulation. Clinical Psychology Review, 101826. doi:10.1016/j.cpr.2020.101826
2. Bae, S., Kang, K. D., Kim, S. W., Shin, Y. J., Nam, J. J., & Han, D. H. (2019). Investigation of an emotion perception test using functional magnetic resonance imaging. Computer Methods and Programs in Biomedicine, 179, 104994. doi:10.1016/j.cmpb.2019.104994
3. Kalantarian, H., Jedoui, K., Washington, P., Tariq, Q., Dunlap, K., Schwartz, J., & Wall, D. P. (2019). Labeling Images with Facial Emotion. Artificial Intelligence in Medicine. doi:10.1016/j.artmed.2019.06.004
4. Sailunaz, K., & Alhajj, R. (2019). Emotion and Sentiment Analysis from Twitter Text. Journal of Computational Science. doi:10.1016/j.jocs.2019.05.009
5. Issa, D., Fatih Demirci, M., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894. doi:10.1016/j.bspc.2020.101894
6. Goldberg, L. R. (1993). The structure of phenotypic personality traits. American Psychologist, 48, 26–34.
7. Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual Review of Psychology, 41, 417–440.
8. 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.
9. Sel, S., & Hanbay, D. (2019). E-Mail Classification Using Natural Language Processing. 2019 27th Signal Processing and Communications Applications Conference (SIU). doi:10.1109/siu.2019.8806593
10. Gupta, S., Agarwal, M., & Jain, S. (2019). Automated Genre Classification of Books Using Machine Learning and Natural Language Processing. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). doi:10.1109/confluence.2019.8776935
11. Akella, R., & Moh, T.-S. (2019). Mood Classification with Lyrics and ConvNets. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). doi:10.1109/icmla.2019.00095
12. Sel, I., Karci, A., & Hanbay, D. (2019). Feature Selection for Text Classification Using Mutual Information. 2019 International Artificial Intelligence and Data Processing Symposium (IDAP). doi:10.1109/idap.2019.8875927
13. M. Aydoğan and A. Karci, "Turkish Text Classification with Machine Learning and Transfer Learning," 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 2019, pp. 1-6.
14. Z. Li, W. Shang and M. Yan, "News text classification model based on topic model," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, 2016, pp. 1-5.
15. Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.
16. https://www.thesaurus.com/, accessed on 12.02.2020
17. Roesslein, Joshua. "Tweepy." Python programming language module (2015).
18. Salton, Gerard, and Donna Harman. Information retrieval. John Wiley and Sons Ltd., 2003.

19. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

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. She has done B.E & M.Tech both in Computer Science & Engineering form VTU Belagavi and qualified in GATE Exam.

Currently she is working on the area of Natural Language Processing, Machine Learning, Digital Image Processing, Data mining, Big-Data Analytics, AI and Pattern Recognition.

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..