Mapping and Visual Analysis of Vocational career with personality trait based on Big Five and Holland's code using Machine Learning

V Karthick, Associate Professor
Department of CSE
Rajalakshmi Engineering College
Chennai, India
vkarthick86@gmail.com

Swathi S, UG Student
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701274@rajalakshmi.edu.in

Tejashree D, UG Student
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701287@rajalakshmi.edu.in

***Abstract*— Working professionals often feel burnt out and lose their interest in the domain they are working under. Sometimes they feel frustrated and work just for survival. The root cause is the lack of choosing the right career path of their interest. Each child at their young age possesses unique traits and habits. As children grow, they exhibit distinct traits and interests. Some are academically inclined, while others harbor passions beyond textbooks. But often they end up choosing the ones forced by their parents and family pressure. Sometimes they don't know which field their interest lies in and finally end up choosing wrong career paths that don't suit them. To address this challenge, this project helps students and children choose their career according to their particular traits and personality. In previous research, the comparison between the ML-augmented method and the traditional profile method in predicting occupational aspiration has been made. The ML-augmented method outperforms the traditional profile method by achieving a higher hit rate of 0.38 and latter having 0.33, thereby marking superior prediction accuracy by ML- augmented method. On the other hand, this paper aims to train data predictive models by using repositories of personality traits using Big Five Personality - and mapping each of the personality to Holland's code vocational career namely RIASEC. By analyzing these models, children can have an idea to pick their career. The integration of K-Means Clustering analyzes personality data to find patterns or groupings that could be useful in psychological research, career counseling, or personal development. Further, each cluster is analyzed using silhouette scores and inertia scores. The mapping of the RIASEC is visualized using heat maps and radar charts.**

Keywords - Holland's Code, RIASEC, Big Five Personality, K-Means, Extraversion, Conscientiousness, Support Vector Machine, Silhouette. Inertia

I. INTRODUCTION

Vocational career prediction stands as anchor to all the career choices made by children and students. This project provides a limelight on the correlation between Holland's code and Big five personality. Holland's code is a theory developed by American psychologist Dr. John L. Holland[3], refers to the taxonomy of interests based on theory of careers and vocational choice Holland's code also referred as RIASEC yields 6 categories of occupations. The RIASEC acronym is short for Realistic, Investigative, Artistic, Social, Conventional. Parallely, the other study - Big Five Personality initially developed by Ernest Tupes, Raymond Christal and, J.M. Digman is a study which is the most commonly adopted model for defining 5 personality traits - Openness, Conscientiousness, Neuroticism, Agreeableness, Extraversion. This paper maps each of the Big 5 personality traits to each RIASEC respectively using Machine Learning

algorithms. This system computes cluster trait scores for models and generates relationships between them. This can help individuals be compatible with a work environment according to their personality.

A mammoth of research has been conducted for thorough analysis of Big Five Personality and Holland's code. [12]et al. Earlier researches were based on Big five personality using Machine learning. One of the research uses focuses on analyzing personality traits using psycholinguistic features extracted from publicly available datasets. It utilizes a set of hand-engineered features, such as LIWC, SenticNet, NRC Emotion Lexicon, VAD Lexicon, and readability measures. Another study identified relationships between congruence and personality in RIASEC types. It mapped the congruence between both models.[11] Another paper is about the 1-D CNN n-grams model proposed by Majumder et al. [3], which achieved the personality prediction performance.

Most of the studies focussed on either the Big 5 personality or RIASEC model separately. Although some studies mapped using psychological factors but did not show visual insights. This grew the motivation behind this paper to be developed. This involves clustering individuals based on their personality surveys and mapping them with RIASEC code, and finally showing visual insights. This study can also help researchers from different fields like psychology, sociology, and data science collaborate to gain deeper insights into human behavior.

II. MATERIALS AND METHODS

DATASET

The dataset was collected from the kaggle site for both Big 5 personality and Holland's code.

Holland's code

The data was hosted on OpenPsychometrics.org, a non-profit effort to educate the public about psychology and to collect data for psychological research. Their notes on the data collected in the codebook.txt. The dataset contains responses, questions, and metadata collected from 145,828 individuals who took the Holland Code (RIASEC) test online. The Holland Codes categorize careers into six types based on personality traits: Realistic , Investigative , Artistic, Social , Enterprising, and Conventional. Participants rated 48 tasks related to these career interests.

Big Five Personality

The dataset contains responses and metadata collected from individuals who took an online personality test based on the "Big-Five Factor Markers." Participants rated statements related to personality traits, such as extraversion, emotional stability, agreeableness, conscientiousness, and openness to experience.

The hardware requirements for the system include an operating system compatible with either Windows or Linux, a minimum of 4 GB of RAM, at least 256 GB of secondary storage, a web camera, and audio speakers. The software requirements for the system include internet browsers such as Chrome, Edge, or Mozilla Firefox, a reliable internet connection, Python version 3.12.3, Jupyter Notebook, and Visual Studio Code installed.

METHODS

K MEANS CLUSTERING

K-Means [20]identifies natural groupings (clusters) within personality data. - We preprocess personality data collected and extract relevant features. K-Means clustering is applied to find distinct personality profiles matching with their suitable work environment. - Each cluster represents a unique combination of personality traits. The goal is to group individuals with similar personality traits into distinct clusters.

1)Data Preprocessing: We start by collecting personality data from individuals

2)Choosing the Number of Clusters (K): Before

applying K-Means, the number of clusters to be created is decided and here we choose 10.

3)Initialization: We randomly select K initial cluster centroids. These centroids serve as the starting points for the clustering process.

4)Assignment Step: Each data point is assigned to the nearest centroid (based on Euclidean distance). This creates K clusters.

5)Clusters: Once convergence is reached, we have K clusters. Each cluster represents a unique combination of personality traits. Individuals within the same cluster share similar personality profiles.

6)Silhouette Score : The silhouette score is a metric used to evaluate the quality of clustering performed by the K-means algorithm.

7)Inertia : Inertia measures the mean squared distance between each data point and the centroid of its assigned cluster. Lower inertia values indicate better clustering because it means that the data points are closer to their respective centroids.

8)Mapping : Now, each trait is mapped with each of the RIASEC codes using the dictionary structure in python.

III. LITERATURE SURVEY

[11, 12] et al. Mehta conducted a comprehensive review of recent advancements in deep learning-based automated personality prediction, emphasizing the effectiveness of multimodal approaches. [11] The study reduced the skewness, and applied a re-sampling technique on the imbalanced dataset. Mairesse et al. developed a feature set for personality prediction. These text features are then fed into traditional machine learning classifiers such as logistic regression, support vector machine (SVM)[17]. Using linguistic cues[10] for the automatic recognition of personality in conversation and text. The authors propose methods for leveraging linguistic features to accurately identify personality traits, advancing the understanding of how language relates to individual characteristics.

Hicks, S.C., R. Liu, Y. Ni, E. Purdom, and D. Risso[5] presents fast clustering for single cell data using mini-batch k-means. This work describes mbk means, a rapid clustering algorithm designed for analyzing single-cell data with mini-batch k-means. Also , the authors show how their approach efficiently clusters big datasets, providing insights into cell population heterogeneity and gene expression patterns.M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf [4] developed support vector machines (SVMs) for classification and regression applications. The authors give an in-depth review of SVMs, covering its theoretical basis, optimization approaches, and practical applications in a variety of disciplines.

A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria use bagged SVM over BERT word embedding ensembles to detect personality traits [7]This work describes a technique for detecting personality traits using bagged Support Vector Machines (SVM) over BERT (Bidirectional Encoder Representations from Transformers) word embedding ensembles. The authors use cutting-edge natural language processing[18] (NLP) techniques to evaluate text data and extract features indicating personality traits. The goal is to improve the accuracy and robustness of personality detection systems by integrating SVM with BERT word embeddings.

[13]Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria discuss recent trends in deep learning-based personality recognition. This review paper explores recent trends in deep learning-based personality recognition. The authors review the many procedures and techniques used in the field, including advances in neural network topologies, training tactics, and feature extraction methods. By combining recent research findings, they highlight emerging trends and future prospects in personality recognition with deep learning algorithms.

More recent work rely on the advances in deep learning and make use of pre-trained word embeddings like Word2Vec [14]and Glove [15] to build better performing personality prediction models. It was found that combining commonsense knowledge with psycho-linguistic features[8] resulted in a remarkable improvement in the accuracy.

[21]The ML-augmented method performs exceptionally well in predicting vocational aspiration. It achieves a significantly higher hit rate (0.44 vs. 0.34) and maintains a lower Euclidean distance (4.23 vs. 4.08) compared to the traditional profile method. However, the profile correlation is slightly lower (0.36 vs. 0.44).

This study looks deep into how our personality traits link with our job choices. It tries to figure out how the unique traits we each have guide us in picking our careers. By looking into real data and breaking down theories, the author highlights what makes us lean toward certain jobs, helping us get the picture of how personality and job interests mix. With this work, we learn more about why we are drawn to specific career paths, thanks to a close look at the ties between our traits and the work we choose[9]

Semeijn, Van der Heijden, and De Beuckelaer (2020) conduct an empirical comparison utilizing the Big Five personality traits to explore the relationship between personality traits and types with career success, providing valuable insights into how individual characteristics impact professional achievement.[16]

In his 2024 doctoral thesis at ВНТУ, the author presents an innovative method for enhancing career advice by combining personality analysis and artificial intelligence. Through merging psychological theories with computational methods, Jaber aims to offer tailored and data-supported suggestions to help individuals navigate their career paths. [6]

[1]El Mrabet and colleagues (2023) have introduced an innovative study in the Proceedings of the International Conference on Smart City Applications. The researchers investigate the utilization of machine learning methods to enhance personality prediction in the educational setting.

[2]Their research delved into the impact of individuals' vocational interests (RIASEC types) and personality traits on their employment outcomes, providing insights into the predictive abilities of these factors in shaping career paths and job roles.

IV. PROPOSED SYSTEM

The proposed work focuses on vocational career[19] prediction based on personality traits. The main aim is to guide individuals, especially young students, in making informed career choices aligned with their interests and traits. The system uses K Means Algorithm to natural groupings (clusters) within personality data. The personality data is preprocessed and relevant features are extracted. This algorithm is applied to find varied clusters and personality profiles that are finally clubbed under the RIASEC code.

Dataset Preprocessing and Cleansing-

- Dataset cleaning and visualization
- Applied K- Means clustering to the features using MiniBatchMeans.
- The number of clusters is set to be 10.

Silhouette score calculation-

S = b-a /max{a,b}

a = Mean distance between the observation and all other data points in the same cluster.

b = Mean distance between the observation and all other data point in the nearest cluster

S = Silhouette score.

Inertia calculation -

$$\sum_{i=1}^{N} (Xi - Ck)^2$$

N is the number of samples

X is the value of each of those samples.

C is the center of the cluster.Personality Trait Score Calculation and Heat Map generation -

- Calculated scores for each of the cluster traits Extroversion, neuroticism, agreeableness, and conscientiousness and normalized.
- Mapped each of the traits with RIASEC types and visualized the compatibility using Heatmap.
- Intensity of color in each heatmap represents the degree of correlation.

Scatterplot Visualization

- The system generates a scatter plot to visualize the relationship between RIASEC types and their corresponding trait scores.
- Each RIASEC type is represented by a point on the plot.

- The x-axis represents the clusters (‘one’, ‘two’, etc.), and the y-axis represents the trait scores.

By examining the scatter plot, one can understand how different clusters align with specific personality traits. For instance, if a cluster is closer to the ‘Realistic’ point, it likely has higher conscientiousness scores.

V. METHODOLOGY

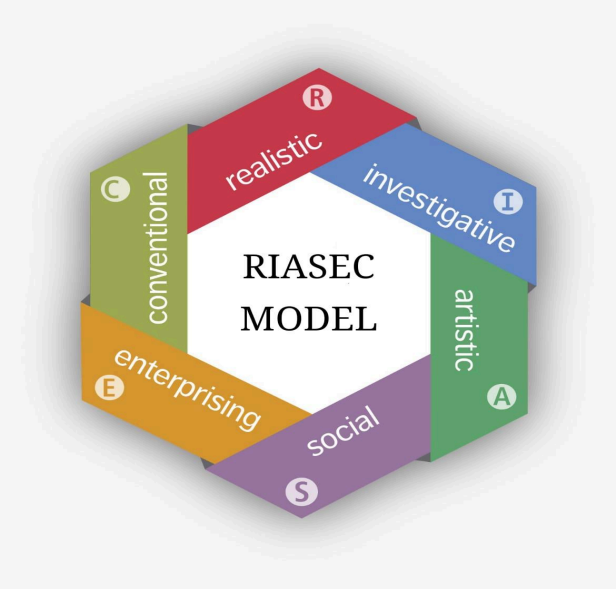

Fig 1. Holland’s code hexagonal model

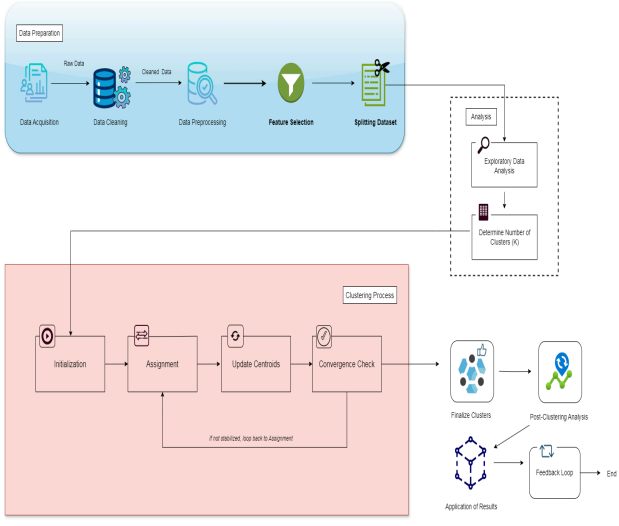

Fig 2. Architecture diagram of RIASEC - Big Five mapping system

VI. RESULTS

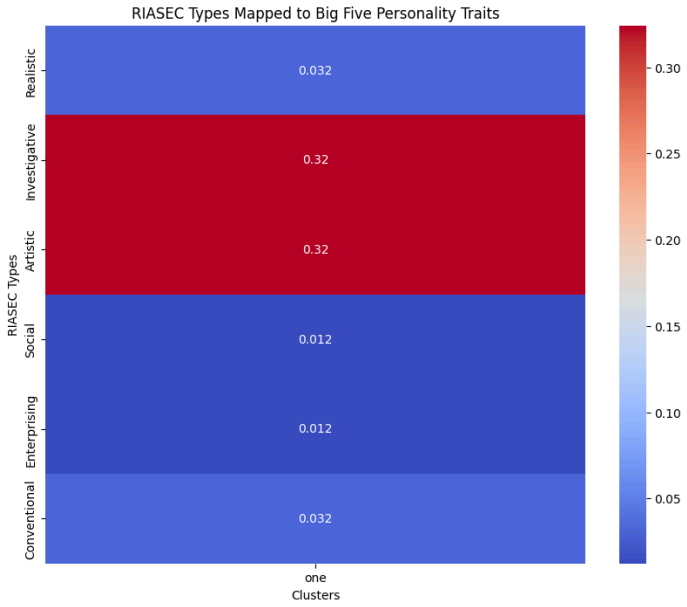

Fig 3. RIASEC Types Mapped to Big Five Personality Trait Scores for cluster 1

Interpretation:

Realistic : This type is positively correlated with Conscientiousness and negatively correlated with Openness.
Investigative : Investigative individuals have a strong positive correlation with Openness.
Social : Social individuals are positively correlated with Extraversion. They are outgoing, friendly, and enjoy social interactions.
Enterprising : This type has a positive correlation with Extroversion.Conventional: Conventional individuals have a strong positive correlation with Conscientiousness.

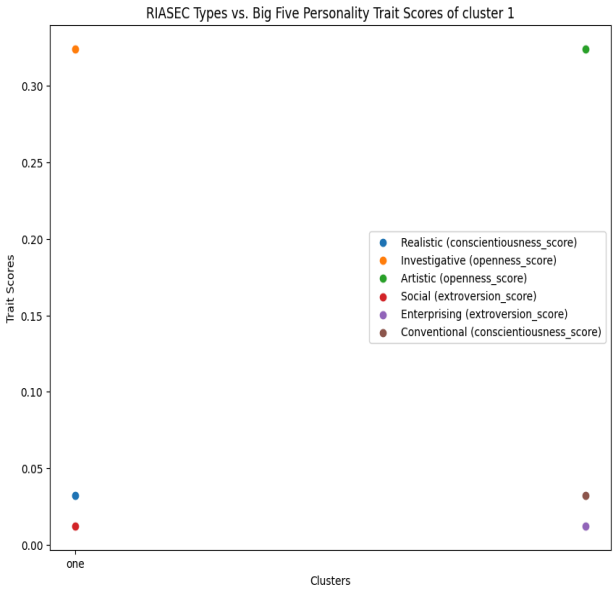

Fig 4. Scatterplot of cluster 1

Interpretation:

1. Realistic: Individuals with this conscientious personality have the highest compatibility with realistic environments.
2. Enterprising: Individuals who are compatible with enterprising environments have lowest conscientiousness. Realistic individuals tend to be more

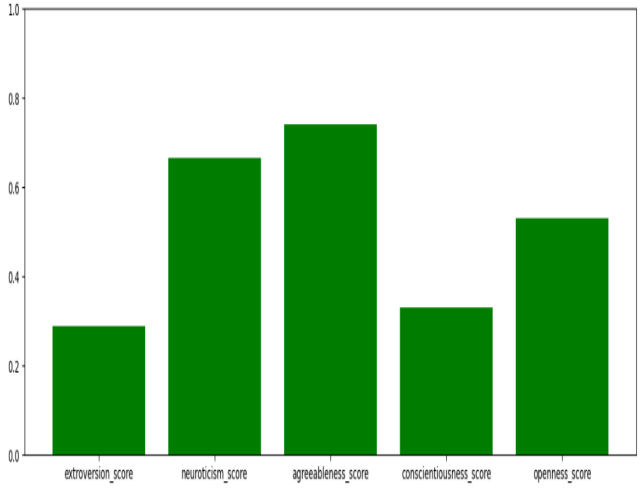

Fig 5.Scores for different personality types for cluster 1

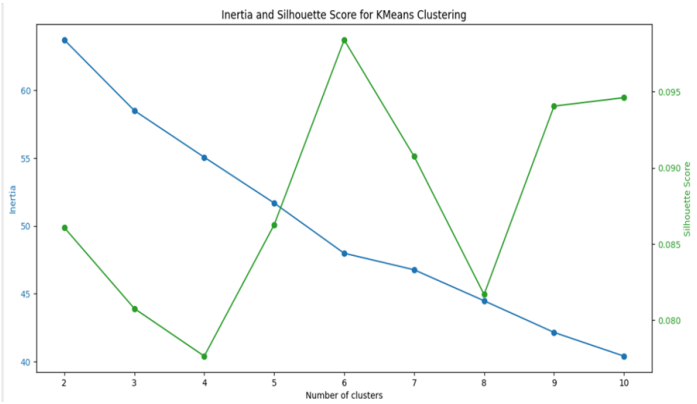

Fig. 6. Analyzing performance of K-Means of inertia and Silhouette Score clustering

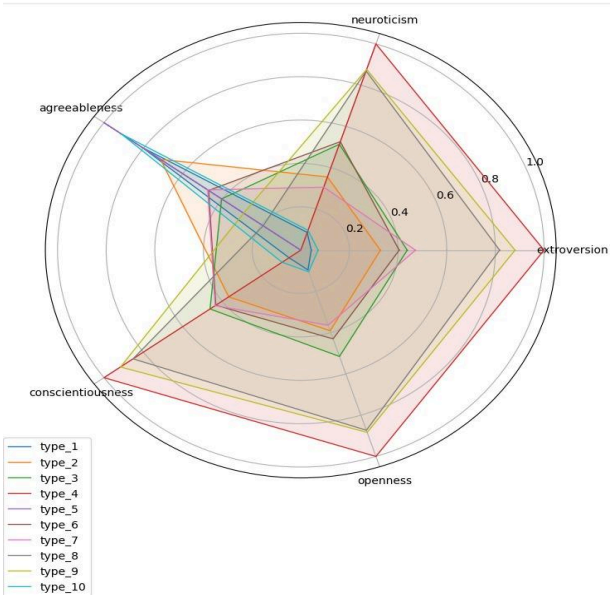

Fig 7. Radar Chart representing to identify strengths and weaknesses of each trait in each cluster

VII. DISCUSSION

In this paper, we have proposed a system by showcasing the effectiveness of machine learning algorithms in predicting the career path of the student. This project shows the key insights into the link between the Big Five traits and Holland's RIASEC career type. This system is trained and data predictive models by using repositories of personality traits

using Big Five Personality - and mapping each of the personality to Holland's code vocational career namely RIASEC. Also, by using K-Means Clustering maps the personality traits with the careers well.The clustering finds the clear personality groups that align closely with the RIASEC Type.

In contrast to previous studies that looked at the Big Five Personality traits and Holland's RIASEC codes separately, this study stands out by combining these two psychological models with the advanced machine learning techniques to provide more accurate insights to choose career paths.

However, this study also possesses some limitations. This study does not generalize all the career contexts. Personality traits are subject to nature and sometimes change over time and the study might be deprecated. Additionally, in technical aspects although K means algorithm might hold the whole game , it may overlook finer nuances within traits. Since most of the datasets for this study rely on self-reporting, it introduces potential biases due to social desirability, and inaccurate self perception.

VIII. CONCLUSION

In this paper, we presented a comprehensive approach to address the challenge of guiding individuals, particularly young students, in making informed career choices aligned with their interests and traits by visualizing their relationship. The project utilized a combination of Holland's RIASEC model and the Big Five Personality traits, using K-means algorithms to predict vocational career paths.

Through the integration of K-Means Clustering, the system clusters within personality data. The accuracy of the clusters is evaluated by 2 metrics namely - 'Inertia' and 'Silhouette. The optimal number of clusters is where the inertia starts to stabilize or decrease at a slower rate. In this case, the inertia decreases from approximately 58 for 2 clusters to around 42 for 10 clusters. The optimal number of clusters could be where the silhouette score is highest. From the table, we can see that it dips at 4 clusters (around 0.084) and peaks around 6 clusters (near 0.092). Based on the results shown below, a reasonable choice might be 6 clusters, where both metrics are relatively favorable.

Number of clusters	Inertia	Silhouette Score
2	~58	~0.095
3	~52	~0.090
4	~48	<0.085
5	>45	>0.085
6	<45	~0.092
7	>45	<0.090
8	<45	>0.087
9	>42	<0.089
10	~42	>0.088

Table 1 : Inertia and silhouette scores for each clusters

By mapping Big Five Personality traits to Holland's code, the system established relationships between personality traits and vocational categories, offering a structured framework for career guidance. The proposed study offers several contributions. Firstly, it provides a data-driven approach to career prediction, moving beyond traditional methods of career counseling. Secondly, it empowers individuals to make informed decisions by aligning career choices with their unique traits and interests. Thirdly, it enhances the effectiveness of vocational guidance.

Future work could focus on further refining the predictive models through larger and more diverse datasets. Additionally, integrating additional factors such as aptitude tests, personal values, and preferences could enhance the accuracy and relevance of career predictions. In addition, creating a system for parents who can record down their childrens' day-to-day activities which in turn can forecast what coursework would interest their children. Overall, the project represents a significant step towards empowering the next generation to make meaningful and fulfilling career choices.

IX. REFERENCES

1. El Mrabet, H., M.A. El Mrabet, K. El Makkaoui, A.A. Moussa, and M. Blej, "Using Machine Learning to Enhance Personality Prediction in Education," *Innovations in Smart Cities Applications Volume 7*, pp. 373–383 (2024).

2. de Fruyt, F., and I. Mervielde, "RIASEC TYPES AND BIG FIVE TRAITS AS PREDICTORS OF EMPLOYMENT STATUS AND NATURE OF EMPLOYMENT," *Personnel psychology*, 52 (3), pp. 701–727 (1999).

3. Gottfredson, G.D., and J.L. Holland, "Dictionary of Holland Occupational Codes," Psychological Assessment Resources Incorporated, (1996).

4. Hearst, M.A., S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE intelligent systems*, 13 (4), pp. 18–28 (1998).

5. Hicks, S.C., R. Liu, Y. Ni, E. Purdom, and D. Risso, "mbkmeans: Fast clustering for single cell data using mini-batch k-means," *PLoS computational biology*, 17 (1),

6. Jaber, A.H., "Enhancing Career Guidance with Personality Insights: A Machine Learning Approach," 2024. https://ir.lib.vntu.edu.ua//handle/123456789/41743 .

7. Kazameini, A., S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality Trait Detection Using Bagged SVM over BERT Word Embedding Ensembles," 2020. http://arxiv.org/abs/2010.01309.

8. Kess, J.F., "Psycholinguistics: Psychology, linguistics, and the study of natural language," John Benjamins Publishing, (1992).

9. López-Muñoz, F., V. Srinivasan, D. de Berardis, C. Álamo, and T.A. Kato, "Melatonin, Neuroprotective Agents and Antidepressant Therapy," Springer, (2016).

10. Mairesse, F., M.A. Walker, M.R. Mehl, and R.K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *Journal of Artificial Intelligence Research*.

11. Majumder, N., S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE intelligent systems*, 32 (2), pp. 74–79 (2017).

12. Mehta, Y., S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features," *2020 IEEE International Conference on Data Mining (ICDM)*, IEEE (2020).

13. Mehta, Y., N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, 53 (4), pp. 2313–2339 (2019).

14. Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013. http://arxiv.org/abs/1301.3781.

15. Poria, S., A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay, "Enhanced SenticNet with affective labels for concept-based opinion mining," *IEEE intelligent systems*, 28 (2), pp. 31–38 (2013).

16. de Raad, B., "The Big Five Personality Factors: The Psycholexical Approach to Personality," Hogrefe Publishing, (2000).

17. Steinwart, I., and A. Christmann, "Support Vector Machines," Springer Science & Business Media, (2008).

18. Wang, H., N. Alanis, L. Haygood, *et al.*, "Using natural language processing in emergency medicine health service research: A systematic review and meta-analysis," *Academic emergency medicine: official journal of the Society for Academic Emergency Medicine*, (2024).

19. Wang, Z., C. Fan, and J. Niu, "Predicting effects of career adaptability and educational identity on the career decision-making of Chinese higher vocational students," *International journal for educational and vocational guidance*, pp. 1–20 (2023).

20. Wu, J., "Advances in K-means Clustering: A Data Mining Thinking," Springer Science & Business Media, (2012).