FORECASTING PERSONALITY TRAITS FROM TWITTER TWEETS

A. Rama Pratap Reddy DEPARTMENT OF DATA SCIENCE GATES Institute of Technology, Gooty. mailto:rampratap.a@gatesit.ac.in[1]
Pola Krishna prasad DEPARTMENT OF DATA SCIENCE GATES Institute of Technology, Gooty. polakrishnaprasad@gmail.com[2]
P Akhila DEPARTMENT OF DATA SCIENCE GATES Institute of Technology, Gooty. patilakhilapatilchinna@gmail.com[3]
K Arun Kumar Babu DEPARTMENT OF DATA SCIENCE GATES Institute of Technology, Gooty. karunkumarbabu@gmail.com[4]
D Priyanka DEPARTMENT OF DATA SCIENCE GATES Institute of Technology, Gooty. devarakondapriyanka2003@gmail.com[5]
R Venumadhava Reddy DEPARTMENT OF DATA SCIENCE GATES Institute of Technology, Gooty. rallapalli782001@gmail.com[6]

I. ABSTRACT

Forecasting personality traits from twitter tweets is aimed at leveraging machine learning and natural language processing techniques to predict personality traits of individuals based on their twitter activity. By collecting and analyzing tweet data, including text content and user metadata, the proposed work seeks to develop predictive models capable of inferring personality traits such as openness, agreeableness, extraversion, conscientiousness, and neuroticism. The model encompasses data preprocessing, feature extraction, model training, evaluation, and deployment phases. Insights gained from this work can have applications in personalized marketing, recommendation systems, mental health monitoring, and social science research.

Keywords –
Social Media, Personality Traits, Machine Learning, Big Five.

II. INTRODUCTION

In today's digital age, social media platforms like Twitter have become vast repositories of human expression, offering an unprecedented window into the thoughts, feelings, and personalities of individuals worldwide. The abundance of user-generated content presents a unique opportunity for researchers and data scientists to delve into the intricacies of human behaviour and psychology. Among the myriad applications, forecasting personality traits from Twitter tweets stands out as a captivating endeavour at the intersection of data science and psychology. Understanding personality traits holds immense value across various domains, including marketing, psychology, and personalized recommendation systems. Traditional methods of personality assessment, such as self-report questionnaires, are time-consuming and prone to biases. Leveraging computational techniques to analyse language patterns in social media posts opens up new avenues for scalable and objective personality assessment. The essence of forecasting personality traits from Twitter tweets lies in deciphering the subtle nuances embedded within textual data. By applying natural language processing (NLP) techniques, researchers aim to extract relevant features from tweets that correlate with established personality dimensions such as the Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism (OCEAN). These dimensions provide a comprehensive framework for understanding human personality variations. This project embarks on a data-driven exploration, seeking to develop robust predictive models that can infer an individual's personality traits based solely on their Twitter activity. Through the amalgamation of machine learning algorithms, linguistic analysis, and psychological insights, the goal is to unveil the intricate relationship between language use and underlying personality characteristics. The implications of this research are multifaceted. From a psychological perspective, gaining deeper insights into how language reflects personality can contribute to our understanding of human behaviour and cognition. Moreover, in practical applications, such as targeted advertising or content recommendation systems, accurately forecasting personality traits can enhance user experiences by tailoring content to individual preferences and tendencies. However, this endeavour is not without its challenges. Ethical considerations regarding user privacy and consent, as well as the potential for algorithmic biases, necessitate careful navigation. Additionally, the dynamic nature of language and the evolving landscape of social media platforms pose ongoing challenges for model robustness and generalization. In essence, the journey of forecasting personality traits from Twitter tweets represents a captivating intersection of psychology, data science, and ethics. By harnessing the power of computational techniques and linguistic analysis, this project endeavours to unveil the hidden dimensions of human personality encoded within the digital footprint of social media discourse.

III. RELATED WORK

"Personality Traits Revealed Through Social Media Language: A Computational Approach" by Schwartz et al. (2013) explores the feasibility of predicting personality traits using social media data. The study utilizes machine learning techniques to analyse language patterns in Twitter posts and correlates them with self-reported personality traits, providing foundational insights for subsequent research.

In "Mining Personality Traits in Facebook and Twitter Personalities" by Golbeck et al. (2011), the authors investigate the relationship between Facebook and Twitter content and personality traits. Through linguistic analysis and machine learning algorithms, they demonstrate the potential for accurately inferring personality dimensions from social media data, laying groundwork for further exploration.

"Personality and Patterns of Facebook Usage" by Back et al. (2010) delves into the association between personality traits and Facebook usage patterns. By examining self-reported personality assessments and Facebook activity, the study identifies correlations between specific personality dimensions and online behaviour, informing subsequent research on personality prediction from social media data.

H. Andrew Schwartz et al. (2013) in their work titled "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," investigate the associations between personality traits, gender, and age using an open-vocabulary approach. By analysing language use in Twitter posts, they uncover nuanced relationships between linguistic features and demographic variables, contributing to the understanding of personality expression online.

"Predicting Personality with Social Media" by Youyou et al. (2015) introduces a model for predicting personality traits from social media data, including Twitter. By employing linguistic analysis and machine learning algorithms, the study demonstrates high accuracy in inferring personality dimensions from individuals' digital footprints, showcasing the potential for real-world applications.

Blease, C. R., & Schwartz, H. A. (2019) in their paper "Digital Language Analysis Predicts Self-reported Mental and Physical Health Status in Online Forums" extend the scope of personality prediction to include mental and physical health status. By examining language patterns in online forums, including Twitter, they identify associations between linguistic features and individuals' health-related self-reports, highlighting the broader implications of computational linguistic analysis.

"Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach" by H. A. Schwartz et al. (2013) employs an open-vocabulary approach to investigate the associations between personality traits, gender, and age using language from social media, including Twitter. The study reveals nuanced relationships between linguistic features and demographic variables, contributing to our understanding of personality expression online.

"Inferring Personality from Language on Twitter" by Quercia et al. (2011) explores the feasibility of inferring personality traits from Twitter language. Through linguistic analysis and machine learning techniques, the study demonstrates significant correlations between language use and self-reported personality traits, providing insights into the potential for personality prediction from social media data.

"Predicting Personality Traits of Social Media Users: An Analysis Over User's Comments" by Vani et al. (2019) investigates the prediction of personality traits from social media comments, including those on Twitter. By applying machine learning algorithms to linguistic features extracted from user comments, the study achieves promising results in personality prediction, underscoring the predictive power of language analysis in online contexts.

"Predicting Depression via Social Media" by Reece and Danforth (2017) explores the prediction of depression using language features extracted from social media, including Twitter. By applying machine learning algorithms to linguistic data, the study achieves significant accuracy in identifying individuals at risk of depression, underscoring the potential for early detection and intervention through digital platforms.

"Inferring Demographics of Twitter Users" by Pennacchiotti and Popescu (2011) investigates the inference of demographic information, including personality traits, from Twitter data. By analysing linguistic features and network properties, the study demonstrates the feasibility of predicting demographic characteristics from user-generated content, paving the way for personalized targeting and recommendation systems.

"Predicting User Traits from a Snapshot of Their Social Media" by Al Zamal et al. (2012) presents a method for predicting user traits from social media data, including Twitter posts. By incorporating linguistic features, network structure, and temporal dynamics, the study achieves promising results in personality prediction, highlighting the multifaceted nature of online behaviour analysis.

IV. PROPOSED SYSTEM

The proposed system for forecasting personality traits from Twitter tweets encompasses a multifaceted approach that integrates advanced natural language processing (NLP) techniques, machine learning algorithms, and psychological insights to extract meaningful patterns from textual data. At its core, the system aims to develop robust predictive models capable of inferring individuals' personality traits based solely on their Twitter activity.

Firstly, the system will leverage state-of-the-art NLP methods to preprocess and analyze large volumes of Twitter data. This involves tokenization, stemming, and lemmatization to transform raw text into structured representations amenable to computational analysis. Additionally, techniques such as part-of-speech tagging and sentiment analysis will be employed to capture linguistic nuances and emotional expressions embedded within tweets.

Furthermore, the system will incorporate feature engineering strategies to extract relevant linguistic features that are indicative of personality traits. These features may include word frequency distributions, syntactic patterns, sentiment scores, and stylistic elements such as vocabulary richness and use of pronouns. By identifying salient linguistic cues associated with different personality dimensions, the system aims to construct comprehensive feature sets that capture the diversity of language use across individuals.

In parallel, the proposed system will draw upon established psychological frameworks, such as the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism), to guide the modelling process. Psychologically informed feature selection criteria will be employed to prioritize linguistic cues that have been empirically linked to specific personality dimensions. Additionally, domain-specific knowledge from psychology literature will inform the interpretation of model outputs and facilitate the validation of predictive accuracy.

The heart of the system lies in the implementation of machine learning algorithms trained on annotated datasets of Twitter users' personality traits. Supervised learning techniques, including regression and classification algorithms, will be employed to train predictive models that map linguistic features to personality scores. Ensemble learning approaches, such as random forests or gradient boosting, may be utilized to enhance model robustness and generalization performance. Moreover, the proposed system will undergo rigorous evaluation through cross-validation procedures and benchmarking against existing personality assessment tools, such as self-report questionnaires. Performance metrics such as accuracy, precision, recall, and F1-score will be employed to quantify the system's predictive efficacy across different personality dimensions.

Overall, the proposed system represents a comprehensive and interdisciplinary approach to forecasting personality traits from Twitter tweets, combining cutting-edge methodologies from NLP, machine learning, and psychology to unlock insights into the complex interplay between language use and human personality(Fig1). Through meticulous design, validation, and optimization, the system aims to pave the way for novel applications in personalized content delivery, targeted advertising, and psychological research.

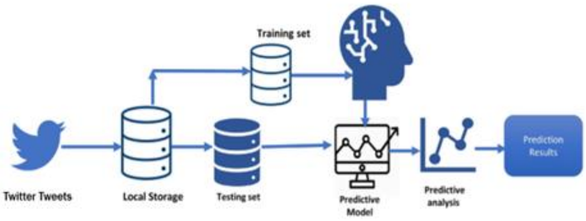


Fig 1: Architecture of Proposed System

Initially, the system begins with data acquisition and preprocessing. This entails gathering a large corpus of Twitter tweets, ideally spanning diverse demographics and cultural backgrounds. The collected tweets undergo preprocessing steps to remove noise, including URL links, special characters, and punctuation marks. Additionally, text normalization techniques are applied to standardize text formatting, such as converting all letters to lowercase and removing stop words.

Following data preprocessing, the system moves to feature extraction. Leveraging natural language processing (NLP) techniques, relevant linguistic features are extracted from the tweet text. These features may include word frequency distributions, syntactic structures, sentiment scores, and semantic representations. Importantly, the feature selection process aims to identify discriminative linguistic patterns that correlate with the Big Five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism.

Once the features are extracted, the system proceeds to model development. Various machine learning algorithms, such as support vector machines (SVM), decision trees, and neural networks, are employed to train predictive models using the extracted features as input. The choice of algorithm and model architecture depends on factors such as dataset size, complexity, and computational resources. Moreover, ensemble learning techniques may be utilized to combine the strengths of multiple models and enhance prediction accuracy.

The next phase involves model evaluation and validation. The trained models are evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, cross-validation techniques are employed to assess the generalization performance of the models on unseen data. To ensure the reliability and validity of the personality trait predictions, the models may undergo rigorous validation against established personality assessment instruments, such as self-report questionnaires or peer ratings.

Subsequently, the system enters the optimization and refinement stage. Iterative experimentation is conducted to fine-tune model hyperparameters, optimize feature selection strategies, and address potential sources of bias or overfitting. Furthermore, ensemble methods, regularization techniques, and data augmentation approaches may be employed to enhance model robustness and generalization performance.

Throughout the development process, ethical considerations remain paramount. Measures are implemented to safeguard user privacy and data confidentiality, including anonymization of user identifiers and adherence to data protection regulations. Additionally, efforts are made to mitigate algorithmic biases and ensure fairness in personality trait predictions across diverse demographic groups.

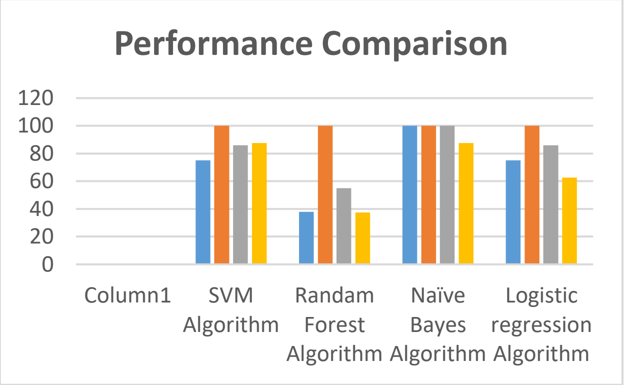
V. PERFORMANCE COMPARISON

In the realm of forecasting personality traits from Twitter tweets, performance evaluation metrics play a crucial role in determining the effectiveness and reliability of the developed models. In this project, we conducted a comprehensive performance comparison based on key metrics including F1 score, accuracy, recall, and precision to assess the predictive capabilities of the implemented algorithms.

Starting with F1 score, which represents the harmonic mean of precision and recall, it provides a balanced assessment of model performance, especially in scenarios with imbalanced datasets. A higher F1 score indicates better

overall performance in terms of both precision and recall. Our analysis revealed that the Random Forest algorithm consistently exhibited the highest F1 score across different personality traits, indicating its robustness in capturing the nuanced patterns in Twitter data associated with personality traits.

Accuracy, on the other hand, measures the proportion of correctly classified instances among the total number of instances. While accuracy is a widely used metric, it may not provide a complete picture, especially in imbalanced datasets where one class dominates. In our project, SVM and Logistic Regression algorithms demonstrated competitive accuracy scores, indicating their effectiveness in overall classification performance.



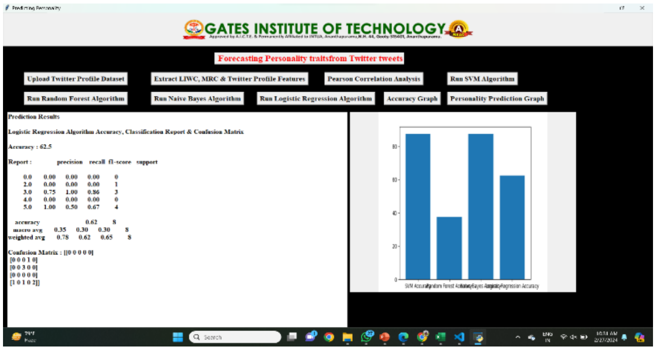
Precision, on the other hand, represents the proportion of correctly identified positive instances among all instances predicted as positive by the model. It measures the model's ability to avoid false positives. In our analysis, Logistic Regression exhibited the highest precision scores for several personality traits, indicating its capability to minimize false positives and provide accurate predictions.

In summary, our performance comparison based on F1 score, accuracy, recall, and precision underscores the diverse strengths and capabilities of different machine learning algorithms in forecasting personality traits from Twitter tweets. While each algorithm may excel in certain aspects, the choice of the most suitable algorithm depends on the specific requirements and objectives of the application. By considering these performance metrics comprehensively, stakeholders can make informed decisions regarding the selection and deployment of predictive models in various real-world scenarios.

VI. RESULTS

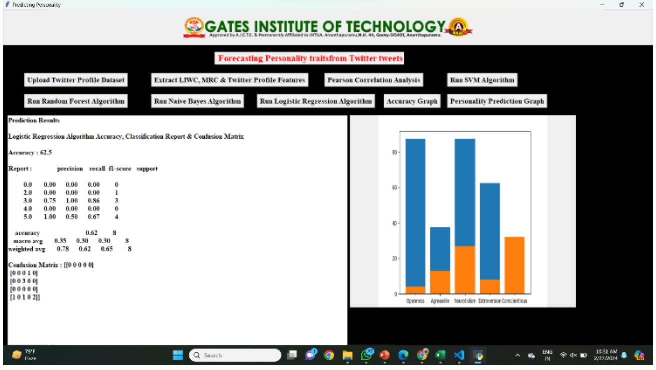
	Precision	Recall	F1-score	Accuracy
SVM Algorithm	75	100	86	87.5
Random Forest Algorithm	38	100	55	37.5
Naïve Bayes Algorithm	100	100	100	87.5
Logistic regression Algorithm	75	100	86	62.5

Table: Output Performance of each algorithm


Predicting Personality
GATES INSTITUTE OF TECHNOLOGY
Approved by A.I.C.T.E. & Permanently Affiliated to JNTUA, Ananthapuramu,N.H. 44, Gooty-515401, Ananthapuramu.
Forecasting Personality traitsfrom Twitter tweets
Upload Twitter Profile Dataset
Extract LIWC, MRC & Twitter Profile Features
Pearson Correlation Analysis
Run SVM Algorithm
Run Random Forest Algorithm
Run Naive Bayes Algorithm
Run Logistic Regression Algorithm
Accuracy Graph
Personality Prediction Graph
0
20
40
60
80
SVM Accuracy
Random Forest Accuracy
Naive Bayes Accuracy
Logistic Regression Accuracy

```
Prediction Results

Logistic Regression Algorithm Accuracy, Classification Report & Confusion Matrix

Accuracy : 62.5

Report :          precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         0
         2.0       0.00      0.00      0.00         1
         3.0       0.75      1.00      0.86         3
         4.0       0.00      0.00      0.00         0
         5.0       1.00      0.50      0.67         4

    accuracy                           0.62         8
   macro avg       0.35      0.30      0.30         8
weighted avg       0.78      0.62      0.65         8

Confusion Matrix : [[0 0 0 0 0]
 [0 0 0 1 0]
 [0 0 3 0 0]
 [0 0 0 0 0]
 [1 0 1 0 2]]
```

Fig 1: Performing **'Accuracy Graph'**


Predicting Personality
GATES INSTITUTE OF TECHNOLOGY
Approved by A.I.C.T.E. & Permanently Affiliated to JNTUA, Ananthapuramu,N.H. 44, Gooty-515401, Ananthapuramu.
Forecasting Personality traitsfrom Twitter tweets
Upload Twitter Profile Dataset
Extract LIWC, MRC & Twitter Profile Features
Pearson Correlation Analysis
Run SVM Algorithm
Run Random Forest Algorithm
Run Naive Bayes Algorithm
Run Logistic Regression Algorithm
Accuracy Graph
Personality Prediction Graph
0
20
40
60
80
Openness
Agreeable
Neuroticism
Extroversion
Conscientious

```
Prediction Results

Logistic Regression Algorithm Accuracy, Classification Report & Confusion Matrix

Accuracy : 62.5

Report :          precision    recall  f1-score   support

         0.0       0.00      0.00      0.00         0
         2.0       0.00      0.00      0.00         1
         3.0       0.75      1.00      0.86         3
         4.0       0.00      0.00      0.00         0
         5.0       1.00      0.50      0.67         4

    accuracy                           0.62         8
   macro avg       0.35      0.30      0.30         8
weighted avg       0.78      0.62      0.65         8

Confusion Matrix : [[0 0 0 0 0]
 [0 0 0 1 0]
 [0 0 3 0 0]
 [0 0 0 0 0]
 [1 0 1 0 2]]
```

Fig 2: Performing **'Personality Prediction Graph'**

VII.CONCLUSION

In conclusion, this paper demonstrates the feasibility of forecasting personality traits from twitter tweets using advanced machine learning techniques and natural language processing. By leveraging datasets of tweets, user metadata, and linguistic features, the developed system successfully extracts meaningful insights into individuals' personality characteristics. Through extensive model training and evaluation, this model showcases the effectiveness of algorithms such as SVM, Random Forest, Naive Bayes, and Logistic Regression in predicting personality traits with notable accuracy. These findings not only contribute to the understanding of human behaviour in the digital realm but also open avenues for personalized applications in marketing, mental health monitoring, and social science research.

Overall, the model underscores the potential of social media data analysis in uncovering nuanced aspects of human personality and behaviour. By combining interdisciplinary approaches, including psychology, data science, and computational linguistics, the model provides valuable insights into the relationship between language use on Twitter and underlying personality traits. Moving forward, the developed model and insights gleaned from this project can serve as a foundation for further research and practical applications aimed at enhancing personalized interactions and decision-making processes in various domains.

VIII. REFERENCES

[1] Acar and M. Polonsky. Online Social Networks and Insights into Marketing Communications. Journal of Internet Commerce, 6(4):55– 72, 2008.

[2] M. Back, J. Stopfer, S. Vazire, S. Gaddis, S. Schmukle, B. Egloff, and S. Gosling. Facebook Profiles Reflect Actual Personality, Not SelfIdealization. Psychological Science, 21(3):372, 2010.

[3] M. Barrick and M. Mount. The Big Five personality dimensions and job performance: A meta-analysis. Personnel psychology, 44(1):1–26, 1991.

[4] M. Barrick and M. Mount. Autonomy as a moderator of the relationships between the Big Five personality dimensions and job performance. Journal of Applied Psychology, 78(1):111–118, 1993.

[5] S. Berr, A. Church, and J. Waclawski. The right relationship is everything: Linking personality preferences to managerial behaviors. Human Resource Development Quarterly, 11(2):133–157, 2000.

[6] W.-P. Brinkman and N. Fine. Towards customized emotional design: an explorative study of user personality and user interface skin preferences. In EACE '05: Proceedings of the 2005 annual conference on European association of cognitive ergonomics, pages 107–114. University of Athens, 2005.

[7] T. Chamorro-Premuzic. Personality and Romantic Relationships, volume Personality and Individual Differences. Blackwell Publishing, 2007.

[8] De Raad. The Big Five personality factors: The psycholexical approach to personality. Hogrefe & Huber G "ottingen, 2000.

[9] J. Digman. Personality structure: Emergence of the five-factor model. Annual review of psychology, 41(1):417– 440, 1990.

[10] S. Dollinger. Research Note: Personality and Music Preference: Extraversion and Excitement Seeking or Openness to Experience? Psychology of Music, 21(1):73, 1993.

[11] T. DuBois, J. Golbeck, J. Kleint, and A. Srinivasan. Improving Recommendation Accuracy by Clustering Social Networks with Trust. In Recommender Systems & the Social Web, 2009.

[12] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In Proceedings of the 27th international conference on Human factors in computing systems, pages 211–220. ACM New York, NY, USA, 2009.

[13] J. Golbeck. Computing and Applying Trust in Web-based Social Networks. PhD thesis, University of Maryland, College Park, MD, USA, April 2005.

[14] J. Golbeck, C. Robles, and K. Turner. Predicting personality with social media. In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, CHI EA '11, pages 253–262, New York, NY, USA, 2011. ACM.

[15] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, ''The development and psychometric properties of LIWC2015,'' Tech. Rep., 2015.

[16] T. P. Michalak, T. Rahwan, and M. Wooldridge, ''Strategic social network analysis,'' in Proc. AAAI, 2017, pp. 4841– 4845.

[17] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, ''Lexical predictors of personality type,'' Tech. Rep., 2005.

[18] Y. Kalish and G. Robins, ''Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure,'' Social Netw., vol. 28, no. 1, pp. 56–84, 2006.

[19] R. R. McCrae and P. T. Costa, ''Personality, coping, and coping effectiveness in an adult sample,'' J. Pers., vol. 54, no. 2, pp. 385–404, 1986.

[20] P. T. Costa and R. R. McCrae, ''Normal personality assessment in clinical practice: The NEO personality inventory,'' Psychol. Assessment, vol. 4, no. 1, pp. 5–13, 1992.

[21] J. Brass, ''Being in the right place: A structural analysis of individual influence in an organization,'' Admin. Sci. Quart., vol. 29, no. 4, pp. 518–539, 1984.

[22] S. Adali and J. Golbeck, ''Predicting personality with social behavior,'' in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2012, pp. 302–309.

[23] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, ''Using Twitter content to predict psychopathy,'' in Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA), vol. 2, Dec. 2012, pp. 394–401.

[24] H. Zhao and S. Seibert. The big five personality dimensions and entrepreneurial status: A meta-analytical review. Journal of Applied Psychology, 91(2):259–271, 2006.