Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

EXPLORING AUTHOR PROFILING FOR PLAGIARISM DETECTION: LEVERAGING PERSONALITY TRAITS AND ENSEMBLE METHODS

VRUSHALI BHUYAR

Department of Computer Applications, Maharashtra Institute of Technology, Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar, India.

SACHIN DESHMUKH

Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar, India.

Abstract

Plagiarism poses a significant challenge in academic circle, as individuals frequently pass off internet content as their own without proper attribution. Traditional detection methods, reliant on established databases, falter when the source material is absent. Author profiling emerges as a crucial tool, analyzing collective language patterns to discern traits like gender, age, native language, and personality. This paper focuses on leveraging personality traits for both plagiarism detection and author profiling. Employing machine learning, particularly ensemble methods, offers promising solutions to these intricate challenges. A dataset of 67 technical research papers, annotated with OCEAN personality traits and plagiarism percentages, underwent preprocessing including outlier detection and normalization. Ensemble techniques, like Extended Gradient Boosting Regressor, Bagging Regressor, Gradient Regressor and AdaBoost Regressor, were applied as base models, with Random Forest Regressor serving as the meta model. Findings reveal notable RMSE values: 0.29 for stacking, 0.93 for averaging, and 2.39 for max voting. Comparison with non-ensemble methods underscores the effectiveness of ensemble learning, notably with Random Forest Regressor achieving a commendable RMSE of 0.29 post-training. Novelty lies in integrating plagiarism detection with personality-based author profiling, providing a comprehensive approach for tackling academic misconduct. By melding machine learning with personality insights, novel avenues for improving detection accuracy emerge. Moreover, ensemble methods enhance the robustness of the approach, showcasing innovative strategies for maintaining academic integrity. This study's findings, integrating plagiarism detection with personality-based author profiling, promise to enhance academic integrity and scholarly conduct, offering valuable insights for refining detection tools and informing decision-making in diverse domains.

Keyword: Plagiarism Detection, Author Profiling, Personality Traits, Machine Learning, Ensemble Methods.

INTRODUCTION

The prevalence of plagiarism [1], [2], [3] poses a significant challenge in academic [4], [5] and professional settings, undermining the integrity of scholarly work and the credibility of authors. Traditional tools [6] of plagiarism detection often rely on matching text against existing databases, which can be limited in cases where the source material is not readily available. To address these shortcomings, there is a growing interest in exploring innovative approaches that go beyond simple text matching. Author profiling [7], [8], which examines the collective language usage patterns of individuals, has emerged as a promising avenue for enhancing plagiarism detection capabilities.

This paper delves into the intersection of author profiling and plagiarism detection, with a specific focus on leveraging personality traits [9], [10] and ensemble methods [11], [12],

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

[13] to bolster detection accuracy. By analyzing not only the textual content but also the underlying characteristics of authors, such openness, consciousness, extraversion, agreeableness and Neuroticism or Emotional Range [14], [15] , we aim to develop a more comprehensive understanding of authorship patterns. Furthermore, by harnessing the power of ensemble methods in machine learning, we seek to enhance the robustness and effectiveness of our detection approach.

In this exploration, we embark on a journey to investigate how personality traits can serve as valuable markers in identifying plagiarized content and attributing authorship. By assembling a diverse dataset of technical research papers annotated with personality traits and plagiarism indicators, we conduct a series of experiments to evaluate the efficacy of ensemble methods in this context. Through our findings, we seek to not only advance the field of plagiarism detection but also shed light on the broader implications of author profiling in understanding and addressing academic misconduct.

This study represents a step towards a more nuanced and sophisticated approach to plagiarism detection, one that takes into account the multifaceted nature of authorship and language usage. As we navigate through the intricacies of author profiling and ensemble methods, we explored innovative strategies for upholding academic integrity and fostering responsible scholarly conduct.

LITERATURE SURVEY

The work of P. Mahbub et al.[16], identified source code segment authorship for various applications like plagiarism detection and digital forensics. Traditional methods struggle with code segments authored by multiple individuals. A recent approach proposes a stacking ensemble classifier integrating deep neural networks, random forests, and SVMs to address this challenge. Compared to single-author methods, this approach shows promising accuracy for multi-authored segments. It utilizes a small set of language-independent code metrics but requires further refinement. Language independence is a strength, though evaluation has focused on Python. Future work could explore metric sets across languages. Challenges include selecting base estimators and fine-tuning multiple classifiers. Overcoming these is crucial for enhancing accuracy in multi-author scenarios.

Muhammad Sajid Maqbool et al. [17], investigated cross-lingual plagiarism detection of digital data in multiple languages. This method proposes using machine learning algorithms, particularly ensemble methods, to detect plagiarism across languages, focusing on the Urdu-English language pair. Evaluation using the CLPD-UE-19 corpus shows promising results, with stacking achieving the highest accuracy at 96%. Detecting plagiarism across languages is challenging due to linguistic differences, but ensemble ML techniques like voting and stacking show superior performance in text classification. Future research could explore leveraging multiple translators and expanding the corpus for further fine-tuning in cross-lingual plagiarism detection.

Eivind Strøm[18] , introduces a solution to the PAN 2021 shared task on style change detection, addressing three sub-tasks through a pragmatic approach involving binary

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

classification and a custom stacking ensemble framework. Results show strong performance across tasks, with the solution outperforming competitors in single- and multi-author classification. While successful, opportunities for further research include identifying additional features and refining multi-label classification techniques.

The work of Todd Zhou et al. [19], applied machine learning to detect cheating in large-scale assessments, mainly focusing on item responses and response times. However, there's been limited exploration of data augmentation within this context. This study fills this gap by investigating data augmentation in the feature space for cheating detection using blending ensemble learning. Four anomaly detection techniques augment the meta-model's input data, improving cheating detection accuracy significantly compared to alternative methods.

M. Gullu et al. [20], focuses on automatically detecting authors from Turkish texts using machine learning. It combines natural language processing for feature extraction, Genetic Algorithms and Bagging for feature selection, and Bagging Algorithm with five classifiers for model creation. Despite a large number of authors, it achieves 89% accuracy, surpassing similar studies. The combination of Genetic Algorithms and Bagging in feature selection enhances accuracy by 8%. Future research could address limitations such as low author diversity and limited text data.

D. Kopev et al. [21], presents a supervised approach to detecting style changes in text documents, achieving success in predicting style changes and identifying their positions. By combining TF.IDF representation with engineered features and employing an ensemble of classifiers, including SVM, Random Forest, AdaBoost, MLP, and LightGBM, the study's approach powered the winning system for the PAN@CLEF 2018 task on Style Change Detection. Recursive application of the model aids in pinpointing exact style breach positions. The study highlights the potential for further improvement through training on varied text lengths, tuning for imbalanced classes, and exploring features based on author-specific linguistic structures.

Stamatatos et al. [22], presents an overview of the PAN/CLEF evaluation lab, a leading platform in text mining research focusing on identifying personal traits of authors in texts. PAN 2015 introduced tasks such as plagiarism detection, author identification, and author profiling, featuring novel approaches like community-driven corpus construction and cross-topic author verification. The lab attracted a large number of global teams, advancing the state-of-the-art in text mining through the development of new corpora and evaluation results. The PAN framework ensures continuity and progress by repeating basic tasks while introducing variations each year, fostering methodological refinement and collaboration among researchers. The requirement for software submissions evaluated within the TIRA platform enhances credibility and reproducibility of results.

Jesus Serrano-Guerrero et al. [23], investigates a growing interest in leveraging social media text for personality recognition, with studies demonstrating the predictive power of linguistic features and deep learning models. Recent research has focused on ensemble methods, combining classic machine learning and deep learning algorithms, to improve accuracy. The proposed stacked ensemble framework integrates diverse algorithms and

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

features, achieving superior performance compared to individual approaches. Future work may explore transformer-based models for enhanced contextual understanding. Overall, the literature underscores the potential of ensemble approaches in personality recognition from social media texts.

Majid Ramezani et al [24], explores how human communication shifts, especially through the internet and social media, impacting personality analysis. It introduces five new methods for Automatic Personality Prediction (APP) from text data, combining them through ensemble modeling to enhance accuracy.

Results indicate the effectiveness of ensemble modeling. The research aims to automate personality analysis, leading to future investigations into the link between personality and various behaviors. It also suggests potential for Automatic Personality Generation (APG) research once APP achieves acceptable performance.

Muntasir Hoq et al. [25], develops an explainable model using programming assignment data to predict final exam grades. SHAP is employed to explain predictions, and mixture models identify student profiles. Results show the model's superiority over other ML methods. Future research includes early prediction using sequential assignment data, investigating problem-solving strategies for assignments of varying difficulties, exploring more complex student profiles, and detecting plagiarism and cheating.

The existing methods for plagiarism detection in academic circles are limited in their effectiveness when the source material is unavailable, leading to a gap in ensuring academic integrity. This study aims to address this gap by leveraging personality traits for both plagiarism detection and author profiling, utilizing machine learning techniques, particularly ensemble methods, to enhance accuracy and robustness.

The play involves employing ensemble machine learning techniques, including Extended Gradient Boosting Regressor, Bagging Regressor, Gradient Regressor, and AdaBoost Regressor as base models, with Random Forest Regressor serving as the meta-model. These models are trained on a dataset consisting of technical research papers annotated with OCEAN personality traits and plagiarism percentages. Preprocessing steps, such as outlier detection and normalization, are performed on the dataset to ensure quality and consistency.

Using ensemble methods, including Gradient Boosting, Bagging, and AdaBoost, with Random Forest as the meta-model, the study trains on a dataset of annotated research papers. By integrating personality traits into the analysis, the approach aims to improve detection accuracy and mitigate academic misconduct.

METHODOLOGY

This manuscript introduces methodology for plagiarism detection using feature extraction based on the OCEAN Big Five personality traits and training the model using a stacking ensemble method with various regression algorithms. Framework for plagiarism detection system using author profiling based on big five personality traits and stacking ensemble method as illustrate in figure 1.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

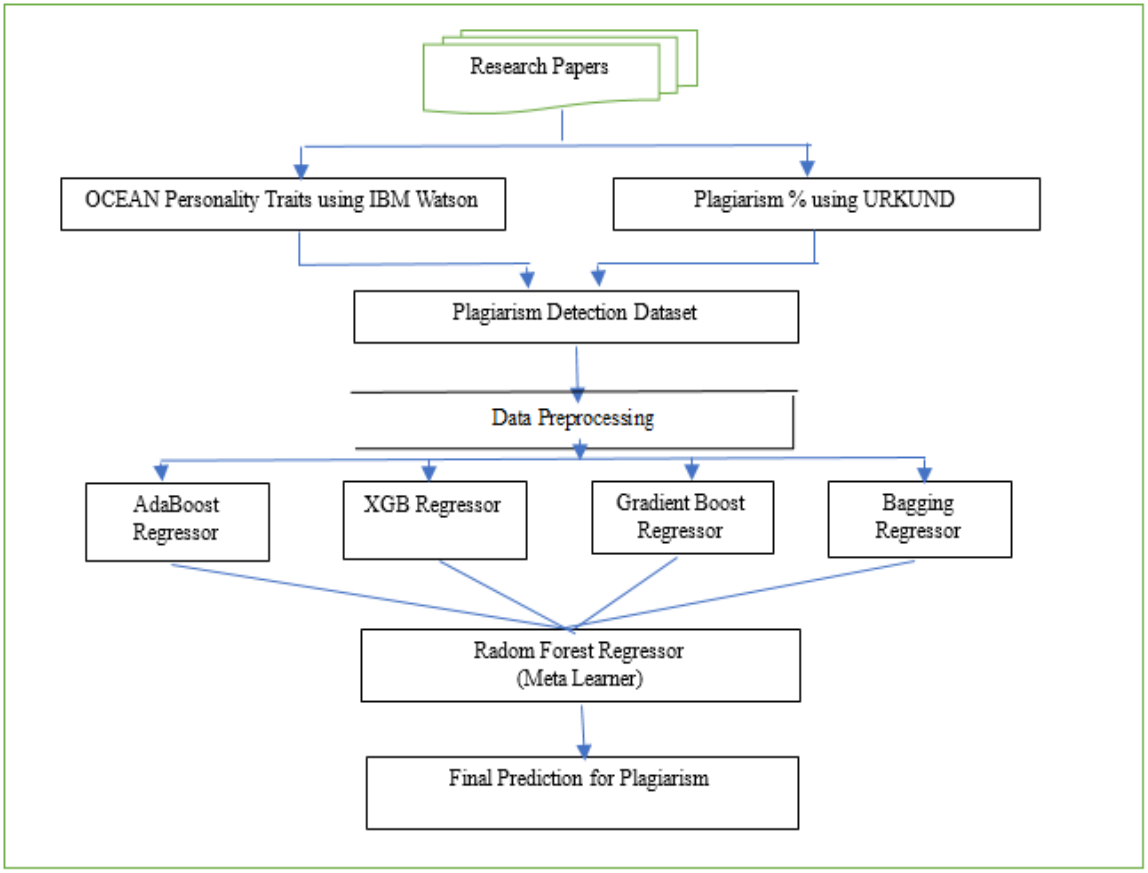

Figure 1: Framework for Plagiarism Detection System using Author Profiling based on Big Five Personality Traits and Stacking Ensemble Method

The study's dataset was constructed by compiling 67 technical papers using URKUND, a popular plagiarism detection software recognized for its accurate similarity assessment. URKUND is widely trusted among scholars for its effectiveness in detecting plagiarism. To extract the Big Five personality traits from these papers, we utilized the IBM Watson Personality Insights API, a renowned tool in the industry known for its reliable personality analysis capabilities. From the IBM Watson API, we gathered five features (OCEAN) along with the plagiarism percentage obtained from URKUND software. Subsequently, the datasets were merged, resulting in a comprehensive dataset consisting of 67 instances, with 40 categorized as plagiarized and 27 as original.

Data preprocessing is a crucial step in refining raw data for analysis by cleansing, transforming, and organizing it to augment its quality. In this study, outlier detection was conducted, and outliers were replaced using the Winsorizer method, which substitutes extreme values with the 95th percentile for upper outliers and the 5th percentile for lower outliers. Subsequently, feature selection was performed by eliminating "PaperName" and "Doc_id," which are unnecessary for subsequent analysis. Following this, Z-score normalization was applied for scaling. With these steps completed, our data is now primed for analysis. Cleaned data consistently yield superior results, underscoring the

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

importance of preprocessing in enhancing machine learning model performance. Ultimately, data preprocessing lays the groundwork to ensure that the data utilized for analysis is dependable and well-suited for its intended purpose.

Stacking Ensemble Method: The stacking ensemble method, also known as stacked generalization, is a powerful technique in machine learning where multiple diverse base models are trained to make predictions on the same dataset. Instead of directly combining the predictions of these base models, a meta-model is trained on the predictions made by the base models. This meta-model learns how to best combine the predictions of the base models, often leading to improved predictive performance compared to any individual base model.

The process typically involves splitting the training data into multiple folds, where each fold is used to train different base models. Then, the base models make predictions on a validation set or the remaining folds. These predictions serve as input features for training the meta-model. Once the meta-model is trained, it can be used to make predictions on new, unseen data. In this study, we employed Extended Gradient Boosting Regressor (XGB), Bagging Regressor (Bagging), Gradient Boost Regressor (GB), and AdaBoost Regressor (AdB) as base models, while Random Forest Regressor (RF) was chosen as the meta-model. Various combinations of algorithms were explored, but ultimately, the aforementioned base models and meta-model yielded superior results in our analysis.

Evaluation Metrics: In the context of detecting plagiarism using author profiling based on Big Five personality traits and employing a stacking regression ensemble method, evaluation metrics play a crucial role in assessing the performance and efficacy of the model. In this study, we have used RMSE.

RMSE measures the square root of the average squared difference between predicted and actual values. In the context of regression-based plagiarism detection using author profiling, RMSE can evaluate the accuracy of predicted plagiarized score based on Big Five personality trait.

Equation 1 represent the formula for Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \ldots (1)$$

Where,

n is number of observations.

y_i represents the actual values.

$\hat{y}_i$ represents the predicted values.

RESULTS AND DISCUSSION

In this study, we compiled a dataset comprising 67 technical research papers with OCEAN factors. The description of the plagiarism detection dataset is presented in Table 1, while Table 2 provides a statistical analysis of the data.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

Table 1: Description of Plagiarism Detection Dataset

S.N.	Features	Description
1	O	Value of Openness as per author
2	C	Value of conscientiousness as per author
3	I/E	Introvert/ Extrovert value
4	Emotional range	Emotional range as per author
5	A	Agreeableness as per author
6	Plagiarism %	Plagiarism Percentage

Table 2: Statistical Description of Data Set

Features	count	mean	std	min	25%	50%
O	67.00	97.73	7.83	40.00	98.00	100.00
C	67.00	46.01	14.08	10.00	36.50	46.00
I_E	67.00	42.30	16.00	7.00	31.00	41.00
Emotional_range	67.00	47.81	28.45	0.00	25.50	49.00
A	67.00	2.03	3.28	0.00	0.00	1.00
Plagiarism%	67.00	21.22	27.78	0.00	3.00	8.00

Ensemble Method: In our dataset, which consists of 67 technical papers and 5 features representing OCEAN traits, we partitioned the data into training and testing sets, utilizing an 80-20% split. Following preprocessing steps, we employed ensemble techniques to enhance our predictive performance. Specifically, we evaluated three ensemble methods: averaging, max voting, and stacking, for comparative analysis.

We incorporated five regression algorithms in our ensemble approach: Random Forest Regressor (RF), AdaBoost Regressor (AdB), Extended Gradient Boosting Regressor (XGB), Gradient Boosting Regressor (GB), and Bagging Regressor (Bagging). All four algorithms served as base models, with Random Forest Regressor employed as the meta-model.

Exploring a total of 26 combinations across these algorithms, our analysis revealed that the stacking method outperformed the averaging and max voting methods, yielding an outstanding RMSE of 0.29. In contrast, the averaging method resulted in an RMSE of 0.93, while the max voting method showed an RMSE of 2.39.

For a comprehensive overview of our findings, please refer to Table 3. Additionally, a graphical representation of the comparison can be found in Figure 2.

Table 3: Results for Averaging, Max Voting and Stacking Ensemble Method

Model	Method	RMSE	Time in seconds
RF+AdB+XGB+GB+Bagging	**Averaging**	**0.93**	**0.12**
RF+AdB+XGB+GB+Bagging	**Max Voting**	**2.39**	**0.11**
RF+AdB	Stacking	0.72	0.13
RF+XGB	Stacking	0.40	0.21
RF+GB	Stacking	0.42	0.12
RF+Bagging	Stacking	0.40	0.20
AdB+XGB	Stacking	0.50	0.18
ADB+GB	Stacking	0.48	0.09
ADB+Bagging	Stacking	0.43	0.09

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

XGB+GB	Stacking	0.47	0.12
XGB+Bagging	Stacking	0.70	0.13
GB+Bagging	Stacking	0.75	0.07
RF+AdB+XGB	Stacking	0.56	0.20
RF+AdB+GB	Stacking	0.49	0.12
RF+AdB+Bagging	Stacking	0.52	0.15
RF+XGB+GB	Stacking	0.31	0.19
RF+XGB+Bagging	Stacking	0.51	0.21
RF+GB+Bagging	Stacking	0.53	0.13
AdB+XGB+GB	Stacking	0.59	0.18
AdB+XGB+Bagging	Stacking	0.58	0.13
AdB+GB+Bagging	Stacking	0.64	0.08
XGB+GB+Bagging	Stacking	0.52	0.15
RF+AdB+XGB+GB	Stacking	0.45	0.19
RF+AdB+XGB+Bagging	Stacking	0.43	0.21
RF+AdB+GB+Bagging	Stacking	0.61	0.15
RF+XGB+GB+Bagging	Stacking	0.67	0.21
ADB+XGB+GB+Bagging	**Stacking**	**0.29**	**0.19**
RF+AdB+XGB+GB+Bagging	Stacking	0.58	0.28

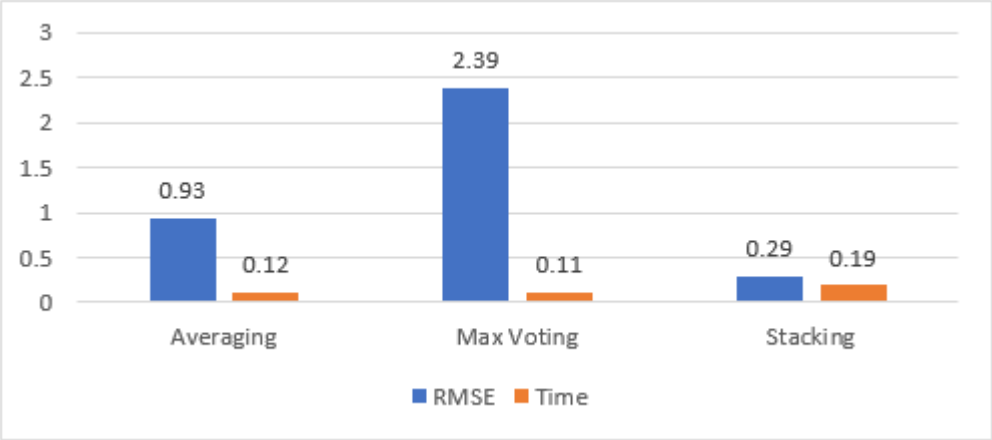

Figure 2: Comparison between Averaging, Max Voting and Stacking Ensemble Method

The results demonstrate that stacking outperforms both max voting and averaging in terms of the root mean square error (RMSE). The exceptional performance of stacking in this scenario highlights its capacity to masterfully blend predictions from base models, navigate data complexity, address model biases, and elevate generalization. These attributes collectively position stacking as a robust ensemble method that frequently surpasses the efficacy of more straightforward techniques like averaging and max voting.

CONCLUSION

This study proposes a novel approach to combating plagiarism by integrating personality-based author profiling with machine learning techniques. By leveraging ensemble methods, such as Random Forest Regressor, Extended Gradient Boosting Regressor,

ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

Bagging Regressor, Gradient Regressor and AdaBoost Regressor, the study achieves significant improvements in detection accuracy. Results indicate that stacking ensemble outperforms traditional methods, with a notable RMSE value of 0.29. This innovative approach promises to enhance academic integrity and scholarly conduct by providing comprehensive detection tools informed by personality insights. The study's findings offer valuable insights for refining detection algorithms and promoting ethical behavior across academic domains.

References

1) Farrokh Habibzadeh, “Plagiarism: A Bird’s Eye View”, J Korean Med Sci. , PP. 38-45, 2023, doi: **10.3346/jkms.2023.38.e373**

2) Ei Mon Phyo, Theoo Lwin, Hpone Pyae Tun, Zaw Zaw Oo, Kyaw Swa Mya & Henry Silverman, “ Knowledge, attitudes, and practices regarding plagiarism of postgraduate students in Myanmar”, Accountability in Research, PP. 672-691, DOI: **10.1080/08989621.2022.2077643**

3) Nur-E Hafsa, “Plagiarism: A Global Phenomenon”, Journal of Education and Practice, Vol.12, PP. 53-59, 2021, DOI: **10.7176/JEP/12-3-08**

4) T. L. Giluk and B. E. Postlethwaite, “Big Five personality and academic dishonesty: A meta-analytic review,” Personality and Individual Differences, vol. 72, pp. 59–67, 2015, doi: **10.1016/j.paid.2014.08.027**.

5) Hloniphani Ndebele, “Demystifying student plagiarism in academic writing: Towards an ‘educational’ solution”, CriSTaL, Vol 8, PP. 39-54, 2020, DOI: **14426/cristal.v8i2.284**

6) Vrushali Bhuyar, S. N. Deshmukh, “Analysis of Support Tools for Plagiarism Detection”, ICAMIDA, ACSR, pp. 38–46, 2023, DOI: **10.2991/978-94-6463-136-4_6**

7) Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, Walter Daelemans, “Overview of the 3rd Author Profiling Task at PAN 2015”, CLEF 2015 Labs and Workshops, Vol 1391,PP. ,2015 http://ceur-ws.org/.

8) D. Radha, Dr. P. Chandra Shekhar, “An Advanced Document Representation Technique Based Approach for Author Profiles Prediction using Word Embedding Techniques”, IJISAE, Vol 12(7s), 377–393, 2023, **https://ijisae.org/index.php/IJISAE/article/view/4081**

9) Feridun Kaya, Fatih Aydin, Astrid Schepman, Paul Rodway, Okan Yetişensoy & Meva Demir Kaya,” The Roles of Personality Traits, AI Anxiety, and Demographic Factors in Attitudes toward Artificial Intelligence” International Journal of Human–Computer Interaction, PP. 497-514, DOI: **10.1080/10447318.2022.2151730**

10) Wiebke Bleidorn , Christopher J. Hopwood , Mitja D. Back , Jaap J. A. Denissen , Marie Hennecke , Patrick L. Hill , Markus Jokela , Christian Kandler , Richard E. Lucas , Maike Luhmann , Ulrich Orth , Brent W. Roberts , Jenny Wagner , Cornelia Wrzus , Johannes Zimmermann, “Personality Trait Stability and Change”, Personality Science, Vol. 2, PP. 1-20, 2021, **https://doi.org/10.5964/ps.6009**

11) K.M.G.S Karunarathna , M.P.R.I.R. Silva , and R.A.H.M. Rupasingha, “Ensemble Learning Approach for Identifying Personality Traits based on Individuals' Behavior”, Sri Lanka Journal of Social Sciences and Humanities Volume 3 Issue 1, 2023, PP. 107-116, **DOI: http://doi.org/10.4038/sljssh.v3i1.91**

12) Muhammad Faraz Manzoor, Muhammad Shoaib Farooq, Muhammad Haseeb , Uzma Farooq , Sohail Khalid , And Adnan Abid, “Exploring the Landscape of Intrinsic Plagiarism Detection: Benchmarks, Techniques, Evolution, and Challenges”, IEEE, Vol 11, 2023, DOI: **10.1109/ACCESS.2023.3338855**

ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 06-2024
DOI: 10.5281/zenodo.12198393

13) Todd Zhou , Hong Jiao, “Exploration of the Stacking Ensemble Machine Learning Algorithm for Cheating Detection in Large-Scale Assessment”, Sage Journals, Vol 83, 2023, **https://doi.org/10.1177/00131644221117193**

14) Peter Eshun, Kyeremeh Tawiah Dabone, Ruth Keziah Annan-Brew, Inuusah Mahama1, Samuel Ofori Danquah, “ Personality Traits and Levels of Self-Efficacy as Predictors of Academic Dishonesty among Higher Education Students in Ghana”, Psychology, 2023, Vol 14, PP. 13-34, DOI: **10.4236/psych.2023.141002**

15) Constantinos M. Kokkinos, Nafsika Antoniadou, Ioanna Voulgaridou, “Majors unleashed: unravelling students’ personality profiles across academic disciplines”, Current Psychology, 2024, **https://doi.org/10.1007/s12144-024-05721-2**

16) Parvez Mahbub, Naz Zarreen Oishie, S M Rafizul Haque, “Authorship Identification of Source Code Segments Written by Multiple Authors Using Stacking Ensemble Method” , ICCIT, Vol 1 , XXXX , 2019, DOI: **10.1109/ICCIT48885.2019.9038412**

17) Muhammad Sajid Maqbool , Israr Hanif, Sajid Iqbal , Abdul Basit and Aiman Shabbir, “Optimized Feature Extraction and Cross-Lingual Text Reuse Detection using Ensemble Machine Learning Models” ,Journal of Computing & Biomedical Informatics, Vol 5, XXXX, 2023, DOI: **https://doi.org/10.56979/501/2023**

18) Eivind Strøm, “Multi-label Style Change Detection by Solving a Binary Classification Problem”, CLEF, Vol 2936, XXXX, 2022,

19) Todd Zhou, Hong Jiao, “Data Augmentation in Machine Learning for Cheating Detection in Large-Scale Assessment: An Illustration with the Blending Ensemble Learning Algorithm”, Psychological Test and Assessment Modeling, Vol 64, PP 425-444, 2022.

20) M. Gullu , H. Polat, “Text authorship identification based on ensemble learning and genetic algorithm combination in Turkish text”, Politeknik Dergisi, Vol 25(3): PP 1287-1297, (2022), DOI: **10.2339/politeknik.992493**

21) D. Kopev, D. Zlatkova, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, and P. Nakov, ‘‘Recursive style breach detection with multifaceted ensemble learning,’’ in Proc. Int. Conf. Artif. Intell., Methodol., Syst., Appl., in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11089, 2018, pp. 126–137, DOI: **10.1007/978-3-319-99344-7_12.**

22) Stamatatos, E, Potthast, M, Rangel, F, Rosso, P, Stein, B, “Overview of the PAN/CLEF 2015 Evaluation Lab” En Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, Proceedings. Springer International Publishing, PP 518-538, 2015 DOI: **10.1007/978-3-319-24027-5_49.**

23) Jesus Serrano-Guerrero, Bashar Alshouha, Mohammad Bani-Doumi, Francisco Chiclana, Francisco P. Romero , Jose A. Olivas, “Combining machine learning algorithms for personality trait prediction” Egyptian Informatics Journal, Vol 25, PP 1-13, 2024, **https://doi.org/10.1016/j.eij.2024.100439**

24) Majid Ramezani, Mohammad-Reza Feizi-Derakhshi,Mohammad-Ali Balafar, Meysam Asgari-Chenaghlu, Ali-Reza Feizi-Derakhshi, Narjes Nikzad-Khasmakhi, Mehrdad Ranjbar-Khadivi, Zoleikha JahanbakhshNagadeh, Elnaz Zafarani-Moattar, Taymaz Rahkar-Farshi, “Automatic Personality Prediction; an Enhanced Method Using Ensemble Modeling”, Neural Computing and Applications, Vol 34, PP 18369-18389, 2022.

25) Muntasir Hoq, Peter Brusilovsky, Bita Akram, “Analysis of an Explainable Student Performance Prediction Model in an Introductory Programming Course”, International Conference on Educational Data Mining, PP. 79-90, 2023, **https://doi.org/10.5281/zenodo.8115693**