

Enhancing plagiarism detection using data pre-processing and machine learning approach

Vrushali Bhuyar¹, Sachin N. Deshmukh²

¹Department of Computer Applications, Maharashtra Institute of Technology, Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar, India

²Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Chh. Sambhajinagar, India

Article Info

Article history:

Received Feb 2, 2024

Revised Nov 19, 2024

Accepted Jan 27, 2025

Keywords:

Data preprocessing

Machine learning

Missing value imputation

Plagiarism detection

Regression

ABSTRACT

Modern technology and the internet have enhanced academic information accessibility, but this has led to a rising global concern about plagiarism. Researchers are actively exploring machine learning as a promising solution for detection. This study underscores the importance of robust data preprocessing for optimal machine learning algorithm performance. Using a dataset of 67 research papers, big five factors (OCEAN), and plagiarism rates, the study employed machine learning to detect plagiarism. The training process involved exposing algorithms to an 80% training subset, followed by evaluating their performance on the remaining 20% in the testing phase, assessing generalization capabilities. For the random forest regressor, bagging regressor, gradient boosting regressor, XGB regressor, and AdaBoost regressor, corresponding root mean squared error (RMSE) are 9.48, 10.66, 11.79, 12.53, and 12.79, respectively. This research contributes novel insights to existing literature by introducing a plagiarism detection model that innovatively integrates outlier detection, normalization, missing value imputation, and feature selection. The unique aspect lies in the effective combination of feature selection and missing value imputation, surpassing previous benchmarks and optimizing precision and efficiency. The approach is metaphorically likened to assembling puzzle pieces, highlighting the distinctive methodology employed in enhancing the performance of the plagiarism detection model using data preprocessing.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Vrushali Bhuyar

Department of Computer Applications, Maharashtra Institute of Technology

Dr. Babasaheb Ambedkar Marathwada University

Chh. Sambhajinagar, India

Email: vrushali.bhuyar@gmail.com

1. INTRODUCTION

In the ever-evolving landscape of academic and digital content, the issue of plagiarism remains a significant concern. As the volume of information available online continues to grow exponentially, traditional methods of detecting and preventing plagiarism are proving to be insufficient. This research aims to address this challenge by proposing an innovative approach that combines advanced data preprocessing techniques with state-of-the-art machine learning algorithms to enhance plagiarism detection.

To extract relevant features from the raw textual data, our approach begins with thorough data preprocessing. To do this, the text must be transformed using methods like missing value imputation, removed unwanted attributes, outlier detection, and normalization into a format that is consistent and appropriate for

machine learning analysis. This thorough preprocessing creates the foundation for machine learning models that are more precise and complex.

The use of machine learning algorithms that can identify patterns and abnormalities in the preprocessed data is the second essential element of our methodology. We train the model on labelled datasets using supervised learning approaches, including regression techniques, so that it can identify patterns linked to plagiarism. Our method seeks to greatly improve the effectiveness of plagiarism detection systems by merging the best features of machine learning and data preprocessing, thereby enhancing the integrity of academic and digital content globally.

Helgesson and Eriksson [1], publishers, and educational institutions must cope with the serious global problem of plagiarism. Even though it's hard to define, we frequently refer to plagiarism as "fictional stealing." It's the same as stealing other people's concepts, credentials, code, and visuals and passing them off as one's own. Because different organizations and fields have varied views on plagiarism, it doesn't need to be reported by a subject expert. It has also been shown to be a ruse in academic [2] and literary circles. They must be found as a result.

Plagiarism has been the subject of numerous studies, which have produced a number of definitions [3]. According to [4] "Plagiarism occurs when someone uses words, ideas, or work products, attributable to another identifiable person or source, without attributing the work to the source from which it was obtained, in a situation in which there is a legitimate expectation of original authorship, in order to obtain some benefit, credit, or gain which need not be monetary". The fast growth of web content has increased the risk of plagiarism in academic research. Plagiarism is a crime in the United States, but violations of the author's right to credit and copyright infringement are also civil wrongs for which the offender may also be punished.

For the purpose of preventing plagiarism and preserving the originality of the information source, it is essential to undertake plagiarism detection at various stages. To do this, research has been going on for a while. Over time, a variety of techniques and technologies have been created to detect plagiarism at various levels [5]–[7]. There are numerous methods for spotting plagiarism; however, it is yet uncertain how these technologies will advance to offer a high level of accuracy.

Machine learning algorithms [8]–[10] play a vital role in scientific computing, as well as in data and text mining. According to recent studies, it is crucial to use data preprocessing and machine learning approaches [11]–[13] to speed up and enhance the performance of the models. As a result, the suggested system relies on machine learning and data preprocessing methods.

Vargas *et al.* [14] presents a comprehensive technique for preprocessing imbalanced datasets in machine learning, addressing the challenge posed by skewed class distributions favoring the majority class. The authors systematically review available methodologies, aiming to provide a comprehensive overview of recent strategies for mitigating class imbalance and improving model performance. Strategies such as cost-sensitive learning, hybrid approaches, oversampling, and under sampling are likely investigated and categorized based on their efficacy with different datasets and algorithms. The survey offers insights into the advantages and disadvantages of each method, furnishing academics and practitioners with a valuable resource for informed decision-making in machine learning projects dealing with imbalanced data.

Alfikri and Purwarianti [15] conduct a thorough literature survey on extrinsic plagiarism detection systems, focusing on machine learning techniques like naive Bayes and SVM. They evaluate diverse literature, emphasizing methods beyond textual analysis, such as metadata examination. Through assessing naive Bayes and SVM across datasets and setups, they provide insights into their effectiveness. The survey not only summarizes existing knowledge but also highlights evolving methodologies and the potential for machine learning to improve detection. By identifying common patterns and unresolved questions, the authors pave the way for further analysis and testing, advancing our understanding of machine learning-based plagiarism detection systems.

Nennuri *et al.* [16] aims to address the growing concern of plagiarism using data mining techniques. The authors present a comprehensive literature survey exploring research in plagiarism detection and data mining. They likely cover various approaches, methodologies, and algorithms used to combat plagiarism in academic and textual content. The study may discuss strengths and drawbacks of different data mining techniques in plagiarism detection, including textual similarity analysis and machine learning. Challenges such as managing paraphrased content and evading detection systems may be highlighted. The survey likely showcases the evolving landscape of data mining technologies for plagiarism detection. Subsequent sections likely discuss the development of an efficient and reliable plagiarism detection system, building upon synthesized research to advance understanding of cutting-edge methods.

Petrides *et al.* [17] investigates the relationships between trait emotional intelligence and the big five personality traits in the Netherlands. The authors investigate how the big five personality traits relate to trait emotional intelligence in a Dutch population. Their study likely reviews existing research on this relationship, considering cultural influences on emotional intelligence and personality traits. Petrides *et al.* [17] work

advances our understanding of individual differences in personality and emotional intelligence within the Netherlands.

Ponomareva *et al.* [18] presents a novel approach for predicting the formation pressure in the Sukharev oil field reservoir in Russia, utilizing multiple regression analysis. The authors develop a prediction model to understand formation pressure dynamics, crucial for reservoir management. Using multiple regression analysis, they consider various variables to improve prediction accuracy. Applied to the Sukharev oil field, the study advances knowledge and offers valuable insights for reservoir engineers. The literature review explores existing methods and sets the context for the authors' innovative approach.

El-Rashidy *et al.* [19] addresses the growing issue of scientific plagiarism using advanced software. It introduces a new database with 42 features for each similarity case, computed using lexical, syntactic, and semantic metrics. The proposed intelligent system, based on long short-term memory (LSTM) architecture, outperforms contemporary systems in detecting text plagiarism on benchmark datasets (PAN 2013 and PAN 2014). The system utilizes innovative methods, including 3D image and signal conversion, preprocessing, and deep learning classification, showcasing superior accuracy in navigating lexical, syntactic, and semantic fluctuations in suspicious cases. Overall, the findings position the proposed system as a leading solution in text plagiarism detection.

Maharana *et al.* [20] underscores the crucial role of data pre-processing in machine learning, addressing challenges like noise, missing data, and inconsistency. It advocates for various pre-processing techniques, emphasizing the importance of cleaning, integration, and transformation to extract meaningful information. The paper introduces data augmentation to handle missing data and enhance model performance. Key points include discussions on image data augmentation techniques to mitigate overfitting, essential data pre-processing steps, and challenges in automating data pretreatment. Guidelines for building models stress the importance of working with correct data, understanding features, and checking distribution. Overall, the findings provide valuable insights for practitioners seeking to optimize machine learning models through effective data pre-processing and augmentation strategies.

Bohra and Barwar [21] explores the impact of deep learning on natural language processing (NLP) and focuses on authorship attribution and plagiarism detection. It introduces a framework involving two layers of processing for authorship attribution, emphasizing the importance of obtaining a similarity score. The proposed plagiarism detection algorithm employs explicit semantic detection and contextualized word embeddings from a BERT pre-trained model. Experiments conducted using Python and the Keras framework demonstrate increased efficiency compared to existing systems, showcasing the algorithm's improved accuracy in identifying plagiarism in NLP task.

Singh and Gupta [22] discusses various extrinsic plagiarism detection techniques, including linguistic, semantic, and contextual approaches, as well as machine learning and deep learning methods. The techniques were implemented on diverse datasets, and their effectiveness was compared using standard measures like precision, recall, F1 Score, PlagDet, and granularity. The study also provides a discussion on the pros and cons of each plagiarism detection technique, offering insights into their strengths and limitations.

2. METHOD

This manuscript introduces a methodology for detecting plagiarism in technical research papers using the big five personality traits (OCEAN). The dataset for this study was assembled by gathering 67 technical papers through URKUND, a widely adopted plagiarism detection software known for its similarity measurement. URKUND is extensively utilized by academics to identify plagiarism effectively [23]. To extract the big five personality factors from technical papers, we employed the IBM Watson personality insights application programming interface (API) [24], [25], a well-regarded tool in the industry known for delivering promising results in personality analysis.

From the IBM Watson API, we collected 21 features, complemented by two additional features-words and plagiarism percentage-extracted from URKUND software. The datasets were then merged, resulting in a comprehensive dataset comprising 67 instances: 40 classified as plagiarized and 27 as unplagiarized. The proposed system involves two key steps: data preprocessing and model building, as illustrated in Figure 1.

2.1. Data preprocessing

Data preprocessing on text documents is essential for cleaning and transforming raw text data into a suitable format for analysis and model training. It helps improve the quality of the data, enhances model performance, and ensures that the resulting insights are meaningful and accurate. Data cleaning involves addressing common issues such as missing values, outliers, and unwanted attributes. Techniques like imputation (filling in missing data), outlier detection (identifying and removing extreme values), and attribute removal (removing irrelevant or redundant features) are essential for ensuring data quality.

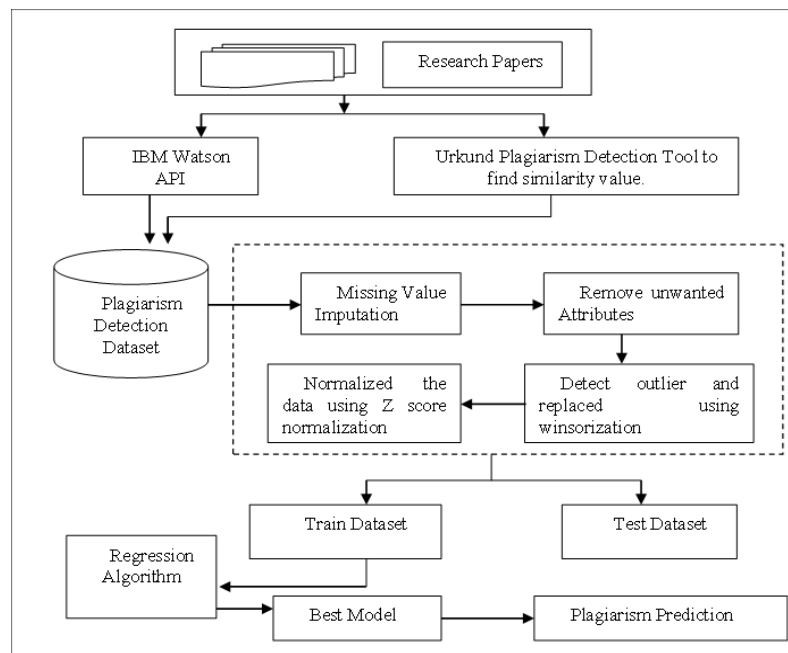


Figure 1. Framework for plagiarism detection system using data pre-processing and machine learning

2.1.1. Missing value imputation

The missing values are indicated for each feature, with varying degrees of missing data. To address the missing values, the framework employs the median imputation method, which replaces the missing values with the median value of the respective feature. This ensures that the dataset is complete and ready for further analysis.

2.1.2. Feature selection

Feature selection is a critical aspect of data preprocessing that improves model performance, reduces overfitting, enhances interpretability, and simplifies the overall analysis process. It is an essential step to ensure that the data used for analysis is relevant, concise, and conducive to building effective and efficient models. Dataset consists of two attributes, "PaperName" and "Doc_id," which are not required for the subsequent analysis and hence removed.

2.1.3. Outlier detection and replaced by winsorization

The next step in the framework involves outlier detection using boxplots. Any outliers identified in the dataset are then replaced using the Winsorizer method. Winsorization replaces extreme values with the 95th percentile for upper outliers and the 5th percentile for lower outliers. By applying this technique, the framework ensures that the dataset is more robust and representative of the data distribution.

2.1.4. Standardization or scaling

Scaling using the Z-score, also known as standardization, is a common preprocessing technique in machine learning to transform numerical features so that they have a mean of 0 and a standard deviation of 1. This process helps in making different features comparable and ensures that the data adheres to a standard normal distribution (mean=0, standard deviation=1). The formula for calculating the Z-score for a particular data point x in a feature is given in (1),

$$Z = (x - \mu) / \sigma \quad (1)$$

Where Z is Z score, X is individual data point, μ is mean of feature and σ is standard deviation of the feature.

2.1.5. Histogram and correlation analysis

Histograms are valuable tools for visualizing data distributions. By grouping data into bins and representing the frequency of occurrences within each bin as a bar, histograms provide a clear visual representation of the data's spread and shape. This visual representation allows for easy identification of key characteristics such as skewness, modality (number of peaks), and the presence of outliers.

Heatmaps are a powerful visualization technique for exploring correlations between features within a dataset. By representing the strength of the relationship between variables using color gradients, heatmaps provide a concise and intuitive visual summary of the data. This visual representation facilitates the identification of potential relationships, dependencies, and redundancies among features, which can be valuable for feature selection, model building, and overall data understanding.

2.2. Model building

The framework applies regression models to the preprocessed dataset for the task of plagiarism detection. Various regression models, including random forest regressor, bagging regressor, gradient boosting regressor, XGB regressor, and AdaBoost regressor, are employed and evaluated based on their adjusted R-squared, R-squared, root mean squared error (RMSE), and the time taken for training. Overall, the framework showcases a comprehensive approach to plagiarism detection, starting from data preprocessing and handling missing values, outliers, and data transformation, to utilizing regression models for prediction. By following this framework, researchers and practitioners can effectively detect instances of plagiarism and enhance the integrity of academic research.

Regression models and sophisticated data preprocessing techniques are combined in a novel way in our suggested method to improve plagiarism detection. Our approach greatly refines the input data by using a multidimensional data preprocessing strategy that includes feature selection and missing value imputation, in contrast to traditional methods that frequently rely on simple preprocessing or low model sophistication. Our method is unique because it combines cutting-edge regression models with this complex preprocessing. Unlike previous research that might concentrate on specific elements or neglect to use state-of-the-art preprocessing methods, our model takes a complete approach to the complexities of plagiarism detection. The combination of sophisticated regression models and careful data preparation produces a stronger system, which is a significant improvement over conventional method.

3. RESULTS AND DISCUSSION

3.1. Dataset creation

In this research, we curated a dataset by gathering 67 technical research papers through URKUND plagiarism detection software. URKUND, a reliable software widely utilized by many academics, was instrumental in this data collection. For feature extraction, we utilized the IBM Watson personality insights API, known for providing robust results and commonly employed by industry professionals. The initial 21 features were extracted from IBM Watson personality traits (O- openness, C- conscientiousness, E- extraversion, A- agreeableness, N- neuroticism or emotional range), while the remaining 2 features were obtained from URKUND software, namely, words and plagiarism percentage. Subsequently, we merged these features to create a dataset comprising 23 distinct features. Description of plagiarism detection dataset is presented in Table 1 and statistical analysis is given in Table 2.

Table 1. Description of plagiarism detection dataset

S.N.	Feature	Description	Missing value
1	O	Value of openness as per author	0
2	C	Value of conscientiousness as per author	0
3	I/E	Introvert/ extrovert value	0
4	Emotional range	Emotional range as per author	0
5	A	Agreeableness as per author	0
6	Curiosity	Value of curiosity as per author	1
7	Practicality	Value of practicality as per author	2
8	Liberty	Liberty value as per author	17
9	Ideal	Value of ideal as per author	50
10	Structure	Value of structure as per author	34
11	Challenge	Value of challenge as per author	38
12	Love	Value of love as per author	46
13	Self expression	Value of self expression	42
14	Harmony	Value of harmony	66
15	Excitement	Value of excitement	49
16	Stability	Value of stability	57
17	Stimulation	Value of stimulation	0
18	Achievement	Value of achievement	0
19	Helping others	Value of helping others	0
20	Taking pleasure in life	Value of taking pleasure in life	0
21	Tradition	Value of tradition	0
22	Words	Value of words	0
23	Plagiarism %	Plagiarism percentage	0

Table 2. Statistical description of data set

	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

3.2. Data preprocessing

This research examined a plagiarism detection dataset, incorporating features related to the author's characteristics, specifically the big five personality traits, and the percentage of plagiarism. A rigorous set of preprocessing steps was employed to ensure the dataset's suitability for subsequent regression model training. To gain insights into the distribution of the data, histograms were generated, offering a visual representation of the spread and frequency of the features. The visualization through histograms provided a distinct portrayal of the data's dispersion and concentration. Figure 2 illustrates this through the depicted histogram.

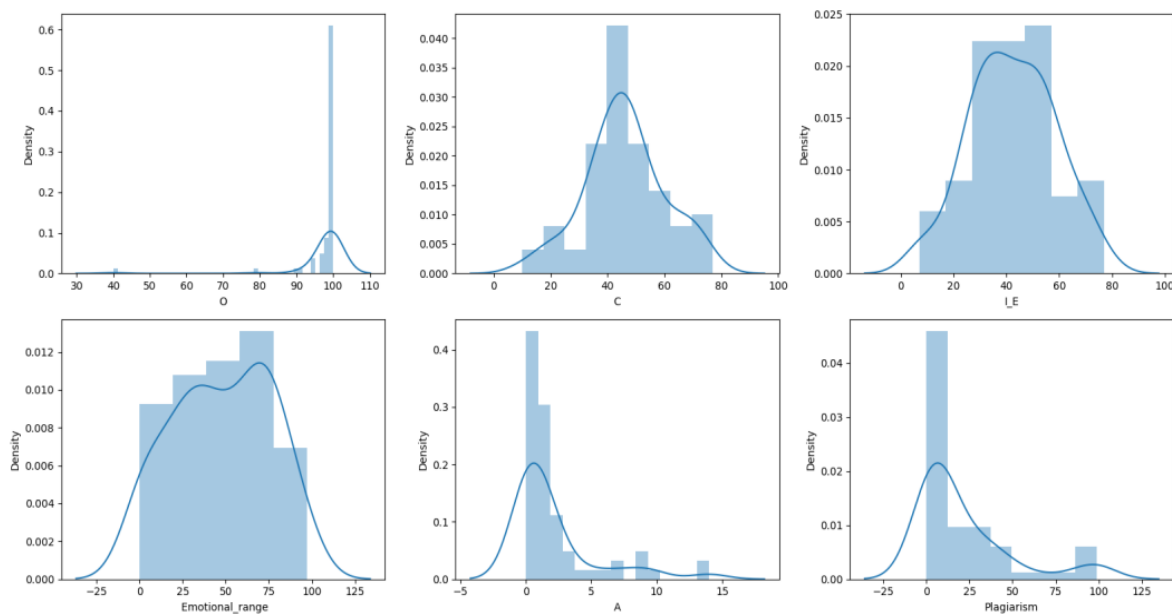


Figure 2. Histogram for openness (O), conscientiousness (c), extraversion (I_E), emotional_range, agreeableness (A), and plagiarism%

3.2.1. Handling missing values

Table 1 reveals a significant number of missing values in the dataset. Both mean and median imputation were considered for filling these gaps. However, mean imputation can be influenced by outliers, leading to less accurate results. To mitigate this issue, we chose median imputation, which is less sensitive to outliers. This approach successfully filled all missing values, creating a complete dataset for further analysis.

3.2.2. Feature selection

Feature selection is crucial step in machine learning and data analysis involving the process of selecting a subset of relevant features for use in model construction. The aim is to improve model performance, reduce computational complexity, and enhance interpretability. Unnecessary attributes, such as "Doc_id," were eliminated from the dataset, streamlining the analysis.

3.2.3. Outlier detection and replaced by winsorization

Utilizing a boxplot for outlier detection (depicted in Figure 3), we identified outliers in the data. Subsequently, we applied winsorization, a technique that replaced upper-side outliers with the 95th percentile of the data and lower-side outliers with the 5th percentile of the data. Following the winsorization process, all outliers were replaced with appropriate values, as illustrated in Figure 4. In both boxplot Figures 3 and 4, the

x-axis represents the combination of personality traits (OCEAN) and plagiarism percentage, while the y-axis depicts the scores or values associated with each combination.

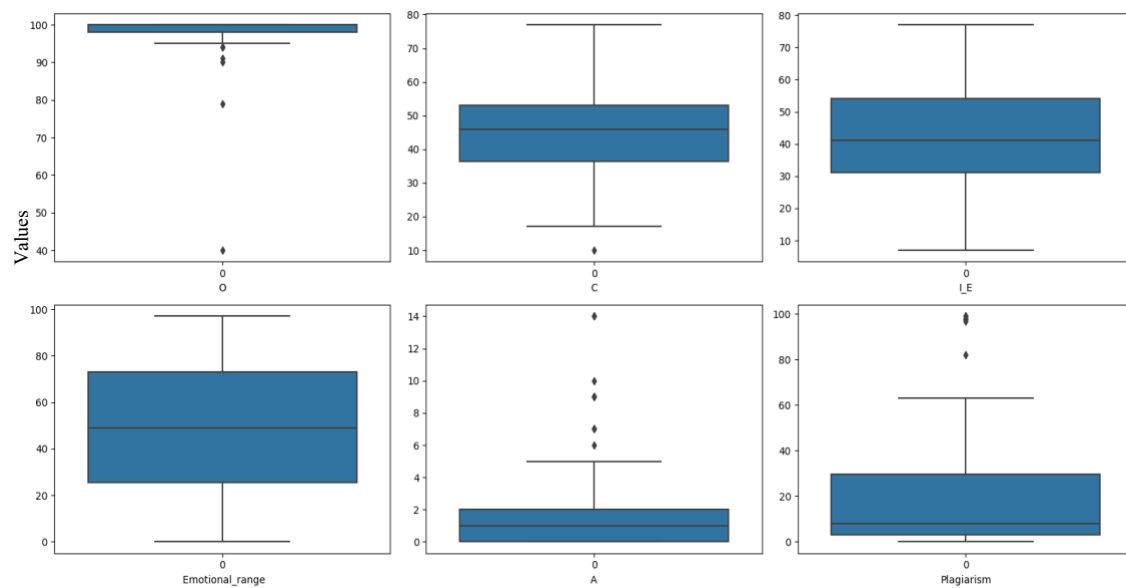


Figure 3. Boxplot with outlier for openness (O), conscientiousness (c), extraversion (I_E), emotional_range, agreeableness (A), and plagiarism%

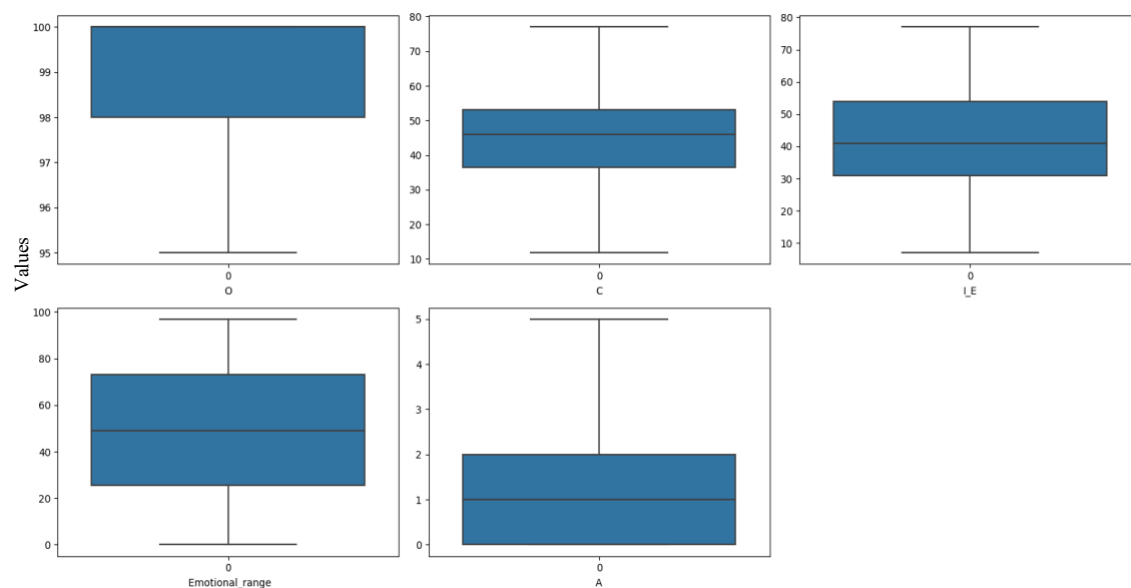


Figure 4. Boxplot without outlier for openness (O), conscientiousness (c), extraversion (I_E), emotional_range, and agreeableness (A)

3.2.4. Standardization or scaling

By employing Z-score normalization to ensure a normal distribution and standardize all attributes. This transformation is essential for many machine learning algorithms. By using Z-score normalization, we converted all values from a 0-100 range to a +1 to -1 range. This standardization helps to prevent features with larger magnitudes from dominating the learning process. The standardized values resulting from this process are displayed in Table 3.

Table 3. After Z score normalization

Sr. No.	O	C	I E	Emotional range	A	Plagiarism
0	0.78	-0.79	0.18	0.82	0.21	5
1	0.78	1.72	0.61	-0.57	0.21	9
2	0.78	0.21	-0.07	-1.35	-1.10	1
3	0.78	0.36	-0.20	-0.53	-1.10	1
4	0.78	1.00	0.18	1.29	0.79	0
...
62	-1.17	0.07	1.42	0.97	0.79	99
63	0.78	-0.72	-0.01	1.08	-1.10	8
64	0.78	-0.15	-0.64	-0.76	-1.10	99
65	0.78	-1.01	-0.71	0.94	-1.10	39
66	-0.51	-0.15	-0.64	-0.42	-1.10	18

3.2.5. Correlation analysis

A heatmap was employed to reveal potential correlations between the big five personality traits and the plagiarism percentage. The analysis of the heatmap indicates that no attribute exhibits a high correlation with another attribute, as their correlation coefficients are all below 0.50. Therefore, all attributes are retained for further analysis. Figure 5 visually represents the heatmap used for correlation analysis.

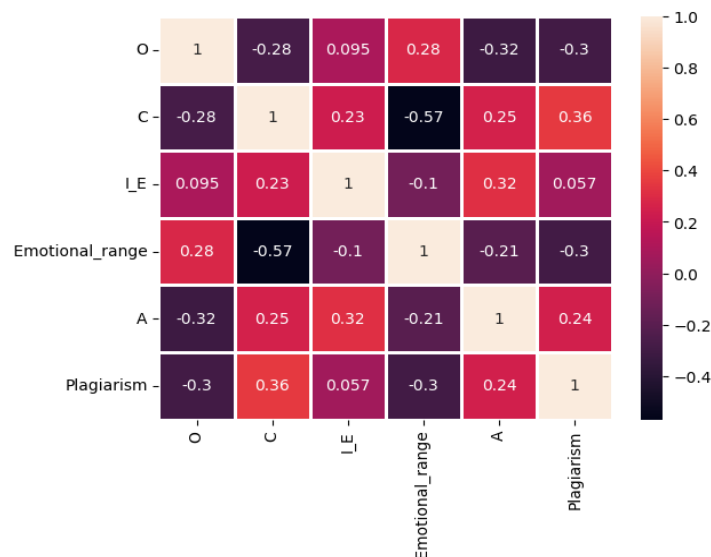


Figure 5. Heatmap for correlation analysis

3.3. Regression models evaluation

Following preprocessing, various regression models underwent training and evaluation utilizing multiple metrics. In this study, we partitioned the dataset into 80% for training and 20% for testing. Lazypredict techniques were employed to identify the top 5 regression models. The regression models were applied to both datasets: one without preprocessing and another with preprocessing. five regression model used which were random-forest regressor, bagging regressor, gradient boosting regressor, XGB regressor, Adaboost regression. Evaluation metrics included adjusted R-squared, R-squared, RMSE, and training time. The performance of the models was scrutinized to assess their predictive capabilities concerning plagiarism percentages based on author traits. Results from Tables 4 and 5 indicate improved performance using preprocessing techniques.

Table 4. Regression model without preprocessing

Model	Adjusted R-Squared	R-squared	RMSE	Time taken
Random Forest Regressor	-0.41	-0.06	34.07	0.05
Bagging Regressor	-0.31	0.01	32.90	0.01
Gradient Boosting Regressor	-0.66	-0.25	37.00	0.01
XGB Regressor	-0.68	-0.26	37.25	0.05
AdaBoost Regressor	-0.36	-0.02	33.48	0.03

Table 5. Regression model with preprocessing

Model	Adjusted R-squared	R-squared	RMSE	Time taken
Random forest regressor	0.76	0.85	9.84	0.13
Bagging regressor	0.72	0.83	10.66	0.03
Gradient boosting regressor	0.66	0.79	11.79	0.03
XGB regressor	0.61	0.76	12.53	0.04
AdaBoost regressor	0.60	0.75	12.79	0.11

This thorough analysis yields valuable insights into preprocessing efficacy, dataset distribution, correlations, and the relative performance of regression models. Figure 6 shows comparative analysis of regression model with and without preprocessing on plagiarism detection dataset. Such insights are crucial for understanding these models' reliability in estimating plagiarism percentages based on author characteristics. The combination of meticulous preprocessing and rigorous model evaluation enhances the findings' reliability and applicability in real-world contexts.

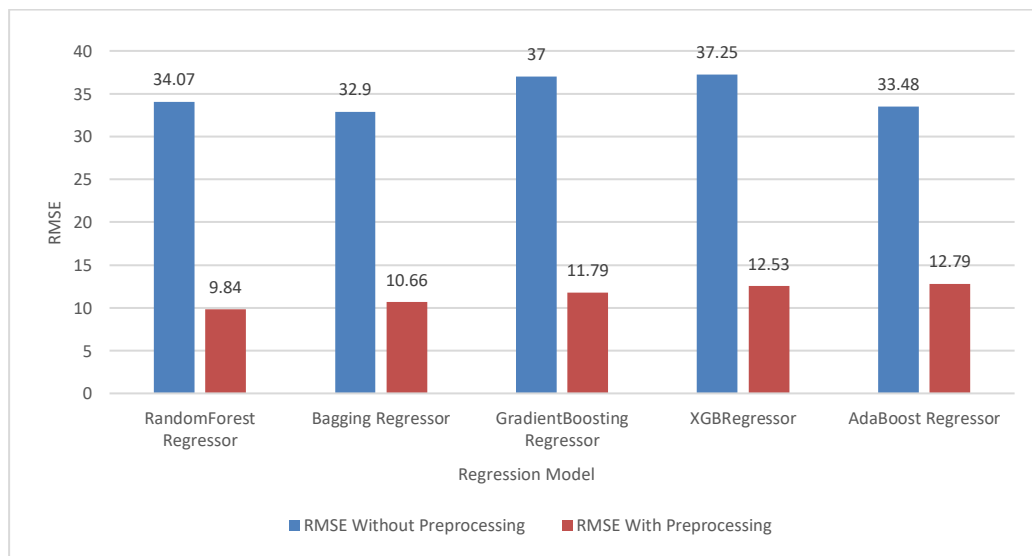


Figure 6. Comparison of regression model with and without preprocessing

4. CONCLUSION

There has been a worrying rise in plagiarism worldwide as a result of the recent explosion in the availability of scholarly content via modern technology and the internet. As a result of researchers' persistent search for workable solutions, machine learning has been investigated as a potential plagiarism detection method. This work highlights the crucial role that data preparation has in optimizing the capabilities of machine learning algorithms and suggests a unique method that combines feature selection and missing value imputation to improve the detection of plagiarism using regression models. Our research is unusual in that it builds a strong architecture for a plagiarism detection model that includes necessary steps including outlier detection, normalization methods, and thorough data pretreatment. Our method adds novel insights to the body of literature by examining the joint effects of feature selection and missing value imputation. Our suggested methodology produced remarkable results after thorough testing on a plagiarism dataset, with RMSE for several regression models of 9.48, 10.66, 11.79, 12.53 and 12.79. Our contribution is essentially the development of a plagiarism detection technology that is incredibly efficient compared to current approaches. The total accuracy and productivity of our plagiarism detection process are maximized by the smooth integration of cutting-edge machine learning techniques with clever data preparation strategies, such as prioritizing important components and filling in information gaps. In conclusion, our method raises the bar for our plagiarism detection tool's performance, much like putting together a puzzle.

FUNDING INFORMATION

No funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Vrushali Bhuyar	✓	✓	✓	✓	✓	✓		✓	✓	✓				✓
Sachin Deshmukh		✓				✓		✓	✓	✓	✓	✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [VAB], upon reasonable request.

REFERENCES




- [1] G. Helgeson and S. Eriksson, "Plagiarism in research," *Medicine, Health Care and Philosophy*, pp. 91–101, 2015, doi: 10.1007/s11019-014-9583-8.
- [2] Q. Gu and J. Brooks, "Beyond the accusation of plagiarism," *System*, vol. 36, no. 3, pp. 337–352, 2008, doi: 10.1016/j.system.2008.01.004.
- [3] H. Maurer, F. Kappe, and B. Zaka, "Plagiarism - a survey," *Journal of Universal Computer Science*, vol. 12, no. 8, pp. 1050–1084, 2017, doi: 10.3217/jucs-012-08-1050.
- [4] T. Fishman, "We know it when we see it is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright," *Fourth Asia Pacific Conference on Educational Integrity (4APCEI)*, 2009, pp. 1-5.
- [5] K. Vani and D. Gupta, "Study on extrinsic text plagiarism detection techniques and tools," *Journal of Engineering Science and Technology Review*, vol. 9, no. 5, pp. 9–23, 2016, doi: 10.25103/jestr.095.02.
- [6] F. Khaled and M. S. H. Al-Tamimi, "Plagiarism detection methods and tools: an overview," *Iraqi Journal of Science*, vol. 62, no. 8, pp. 2771–2783, 2021, doi: 10.24996/ijcs.2021.62.8.30.
- [7] M. N. Mansoor and M. S. H. Al-Tamimi, "Computer-based plagiarism detection techniques: A comparative study," *International Journal of Nonlinear Analysis and Applications*, vol. 13, no. 1, pp. 3599–3611, 2022, doi: 10.22075/ijnaa.2022.6140.
- [8] W. Anwar, I. S. Bajwa, and S. Ramzan, "Design and implementation of a machine learning-based authorship identification model," *Scientific Programming*, vol. 2019, no. 1, 2019, doi: 10.1155/2019/9431073.
- [9] J. Burdack, F. Horst, S. Giesselbach, I. Hassan, S. Daffner, and W. I. Schöllhorn, "Systematic comparison of the influence of different data preprocessing methods on the performance of gait classifications using machine learning," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020, doi: 10.3389/fbioe.2020.00260.
- [10] M. M. Zahid, K. Abid, A. Rehman, M. Fuzail, and N. Aslam, "An efficient machine learning approach for plagiarism detection in text documents," *Journal of Computing & Biomedical Informatics*, vol. 4, no. 2, pp. 241–248, 2023.
- [11] T. Foltýnek, N. Meuschke, and B. Gipp, "Academic plagiarism detection: a systematic literature review," *ACM Computing Surveys*, vol. 52, no. 6, pp. 1–42, 2019, doi: 10.1145/3345317.
- [12] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine*, vol. 220, pp. 1–12, 2022, doi: 10.1016/j.cmpb.2022.106773.
- [13] M. A. El-Rashidy, R. G. Mohamed, N. A. El-Fishawy, and M. A. Shouman, "An effective text plagiarism detection system based on feature selection and SVM techniques," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 2609–2646, 2024, doi: 10.1007/s11042-023-15703-4.
- [14] V. W. D. Vargas, J. A. S. Aranda, R. D. S. Costa, P. R. S. Pereira, and J. L. V. Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowledge and Information Systems*, vol. 65, no. 1, pp. 31–57, 2023, doi: 10.1007/s10115-022-01772-8.
- [15] Z. F. Alfikri and A. Purwarianti, "Detailed analysis of extrinsic plagiarism detection system using machine learning approach (naive Bayes and SVM)," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 12, no. 11, 2014, doi: 10.11591/telkomnika.v12i11.6652.
- [16] R. Nennuri, M. G. Yadav, M. Samhitha, S. S. Kumar, and G. Roshini, "Plagiarism detection through data mining techniques," *Journal of Physics: Conference Series*, vol. 1979, no. 1, 2021, doi: 10.1088/1742-6596/1979/1/012070.
- [17] K. V. Petrides, P. A. Vernon, J. A. Schermer, L. Lighthart, D. I. Boomsma, and L. Veselka, "Relationships between trait emotional intelligence and the big five in the Netherlands," *Personality and Individual Differences*, vol. 48, no. 8, pp. 906–910, 2010, doi: 10.1016/j.paid.2010.02.019.

Enhancing plagiarism detection using data pre-processing and machine learning ... (Vrushali Bhuyar)




- [18] I. N. Ponomareva, D. A. Martyushev, and S. K. Govindarajan, "A new approach to predict the formation pressure using multiple regression analysis: case study from Sukharev oil field reservoir – Russia," *Journal of King Saud University - Engineering Sciences*, vol. 36, no. 8, pp. 694-700, 2022, doi: 10.1016/j.jksues.2022.03.005.
- [19] M. A. El-Rashidy, R. G. Mohamed, N. A. El-Fishawy, and M. A. Shouman, "Reliable plagiarism detection system based on deep learning approaches," *Neural Computing and Applications*, vol. 34, pp. 18837–18858, 2022, doi: 10.1007/s00521-022-07486-w.
- [20] K. Maharana, S. Mondal, and B. Nemade, "A review: data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.
- [21] A. Bohra and N. C. Barwar, "A deep learning approach for plagiarism detection system using BERT," *Congress on Intelligent Systems*, pp. 163–174, 2022, doi: 10.1007/978-981-16-9113-3_13.
- [22] M. Singh and V. Gupta, "Review of extrinsic plagiarism detection techniques and their efficiency comparison," *Advanced Network Technologies and Intelligent Computing*, pp. 609–624, 2022, doi: 10.1007/978-3-030-96040-7_46.
- [23] V. Chandere, S. Satish, and R. Lakshminarayanan, "Online plagiarism detection tools in the digital age: a review," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 1, pp. 7110–7119, 2021.
- [24] R. A. P. Junior and D. Inkpen, "Using cognitive computing to get insights on personality traits from twitter messages," *Advances in Artificial Intelligence: 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017*, pp. 278–283, 2017, doi: 10.1007/978-3-319-57351-9_32.
- [25] Z. Balogh, "Analysis of public data on social networks with IBM Watson," *Acta Informatica Malaysia*, vol. 2, no. 1, pp. 10–11, 2018, doi: 10.26480/aim.01.2018.10.11.

BIOGRAPHIES OF AUTHORS



Vrushali Bhuyar    is currently working as an Assistant Professor, Department of Computer Applications, Maharashtra Institute of Technology, Chh. Sambhajinagar (MS), India. She has more than 15 years of teaching experience for post graduate (MCA). She worked as BOS member for MCA. She has published research articles in reputed international journals and conferences. Her research interest is data mining, text mining, machine learning, and data science. She can be contacted at email: vrushali.bhuyar@gmail.com.



Dr. Sachin N. Deshmukh    is currently working as a Professor, Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, (MS), India. He is also working as a Director, Innovation, Incubation and Linkages Center. Currently he is a member of Purchase Committee, Senate and Academic Council of the University. In initial phase, he was a member of Advisory Board NPIU-MERITE, MHRD, Government of India. He was a member of Management Council of the University. He has more than twenty-eight years of experience in teaching for post graduate (M. Tech., M.Sc., and MCA) and graduate courses (B.E. /B.Tech.) and also have an experience of software development. He has given expert talks in the colleges and universities on different recent technologies. He has published more than 100 research articles in reputed international journals and conferences and 9 students has been completed Ph.D. under his guidance and 6 are working. He can be contacted at email: sndeshmukh@hotmail.com.