International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Stylometric Feature Extraction and Machine Learning Classification for Research Paper Plagiarism Detection

Dr.M.AyyappaChakravarthi[1], K.L.Roshini[2], K.S.R.S.Vaishnavi[3], Ch.Aparna[4], E.SaiVarshika[5]
[1]Associate Professor, Dept. of CSE–Data Science
KKR & KSR Institute of Technology and Sciences, Guntur
[1]dr.chakravarthy.cseds@kitsguntur.ac.in
[2,3,4,5]B.Tech Student, Dept. of CSE–Data Science
KKR & KSR Institute of Technology and Sciences, Guntur
lakshmiroshinikotha@gmail.com[2], srivaishnavikompalli@gmail.com[3], aparnachikkala08@gmail.com[4],
enukollusaivarshika89@gmail.com[5]

***Abstract*—Plagiarism in research papers has become hard to detect due to the rise in the use of online materials and the development of artificial intelligence-assisted writing tools. Traditional plagiarism detection techniques through text matching tend to perform poorly, especially if the plagiarized text is paraphrased, partially plagiarized, or rewritten in different styles. This paper proposes a plagiarism detection methodology focusing on stylometric analysis. Instead of relying on text matching, the methodology utilizes machine learning classification based on feature extraction. The methodology analyzes the research paper from word, syntax, and structure perspectives. to grasp the distinct writing patterns of the research paper author and pinpoint irregular patterns indicative of plagiarized or foreign text. Utilized features include word frequency, Part of Speech, punctuation, and sentence structure, which are measured and normalized as vectors. These vectors help train traditional classifiers such as K-Nearest Neighbors, Naive Bayes, Decision Trees, ensemble models like XGBoost, and stacking classifiers. These trained classifiers have the ability to distinguish between author style and suspicious pieces of content with high precision. The tool provides both document-level and section-level features and is capable of identifying fully plagiarized documents, as well as hybrid documents that contain only some sections of suspicious content. This tool assists supervisors, reviewers, and institutions in maintaining academic integrity because of the in-depth style insights that lie beyond mere similarity scanning.**

***Index Terms*—Research Paper Plagiarism Detection, Stylometric Feature Extraction, Machine Learning Classification, AI- Generated Text Detection, Authorship Attribution, Academic Integrity, Writing Style Analysis, Ensemble Learning, XGBoost Classifier, Text Classification**

I. INTRODUCTION

Plagiarism in academic writing affects the integrity, validity, and quality of scientific research. [4] The majority of current solutions for plagiarism detection rely on string matching and document similarity analysis, which fail if plagiarism occurs through rewritten text and copying from other sources. The emergence of AI-based writing assistants has resulted in more effective generation of text that meets requirements for avoiding copying but violates principles of originality in academic work. The challenge lies in having tools whose focus is on how a text is written rather than what is written in a text. Stylometry refers to analysis of writing styles by using quantitative linguistic features. By building a model of a person's usual writing style and matching it to other parts of a text, it is feasible to identify parts of a text which didn't result from a given author's effort. It has been observed that stylometry has been able to differentiate between texts written by a person and texts written by machines as well as texts written by different people in the same text, a fact which makes stylometry suitable for sophisticated techniques of plagiarism detection in academic texts.

The proposed project combines the extraction of stylometric features and machine learning algorithms to detect the suspicious regions in research papers. The concept is developed to identify papers written by the same author and those papers which may comprise sections of plagiarism, rewritten, or outsourced content. This technique helps to develop a smarter way of supporting research by concentrating on the variations in writing style, and not the text. Additionally, the system can adjust to various writing styles and changing types of academic misconduct thanks to the integration of stylometric analysis with machine learning. The model can learn subtle stylistic deviations that are challenging to identify through surface- level comparison by training classifiers on meticulously labeled datasets that contain authentic, plagiarized, and AI- assisted texts. This method provides comprehensive insights into possible

II. PROBLEM STATEMENT

Plagiarism in research papers is a rising issue that affects academic integrity. Traditional plagiarism detection techniques are primarily based on text similarity and are only useful for detecting direct plagiarism. These techniques are not very useful in detecting intelligent plagiarism, like paraphrasing and style changes.

Stylometric analysis, which involves the study of writing styles using linguistic and statistical properties, offers a possible remedy. By integrating stylometric feature extraction with machine learning classification, plagiarism can be identified through inconsistencies in writing style rather than text similarity.

However, the current methods have some difficulties in feature selection, scalability, and accuracy for different academic texts. Thus, a plagiarism detection system based on stylometric features and machine learning is needed to accurately detect direct and intelligent plagiarism in research papers.

A system like this can precisely identify stylistic deviations within a single manuscript by analyzing research papers at various granularities, such as document, section, and paragraph. Machine learning models can discern between authentic variations and suspicious anomalies resulting from paraphrased, outsourced, or AI-generated content by learning an author's typical writing patterns from trustworthy training data. This method lessens reliance on external source repositories while simultaneously improving detection accuracy. As a result, a stylometry-based plagiarism detection framework can be a useful supplement to conventional similarity-based techniques, providing a more reliable and up-to-date way to maintain originality and credibility in scholarly work.

III. Objectives

To develop an automated system for plagiarism detection of research papers through stylometric analysis.

Stylometric analysis: As a special application based on statistical analysis, the purpose of stylometric analysis is primarily to obtain stylometric features such

For the classification of research documents as original and plagiarized through supervised machine learning algorithms.

To address the limitations of conventional string matching tools to identify paraphrased and obfuscated information.

To compare the accuracy and efficiency rates of various classifiers to distinguish between authorship patterns

This study will also explore whether machine learning models based on stylometric features are capable of picking up distinctive authorship patterns in academic texts. Based on lexical, syntactic, and structural feature analysis, the paper examines various supervised classifiers with respect to their performance in distinguishing between original contents and plagiarized or stylistically inconsistent text. An emphasis is laid on overcoming the weakness of a traditional string-matching approach, mainly regarding the detection of paraphrased, obfuscated, and AI-assisted writing. A comparison of several classifiers allows this work to determine the most accurate and efficient models for intrinsic plagiarism detection and provide practical insights into the development of robust, automated systems of plagiarism detection.

IV. Literature Review

Stylometric analysis is also used extensively in authorship attribution, where it is required to authenticate or find out the author of any literary work based on style.[1]. They include word frequencies, lexical variation, sentence lengths, punctuation marks, and part-of-speech tagging distributions. These characteristics are used to mark typical styles, which are very hard even to forge, even if it is about other topics and even if new vocabularies are used. Previous research works have employed stylometry on various aspects, including authorship, security, and AI-based literary forgery, showing good results in their accuracy rates in classification problems. Recent research shows that by combining stylometric variables and machine learning algorithms, good distinction between human writings and AI scan be attained.[2]Toboostaccuracy, various machine learning algorithms such as 3 XGBoost and stacking have been used successfully by training on vectors distinct for writing style features.[3] Classical algorithms such as k-Nearest Neighbor, Naive Bayes, and Decision Tree have also been utilized successfully as benchmarks by carefully setting their differing parameters.[5] These experiments prove that style-representations can be on an equal footing with and even outperform TF-IDF-style representations. [6]Concerning plagiarism detection, two types of solutions have been rec-ognized as more prominent the extrinsic approach and the intrinsic approach. For the former, the similarity of the suspi-cious document is sought with other documents from external sources, whereas intrinsic plagiarism identification focuses on the self-coherence of style within the document under exami-nation and concentrates on style-switches potentially showing outsourced sections of text. [7] Research regarding intrinsic plagiarism detection indicates that style representation on the paragraph or section level scan assist with the identification of questionable text fragments even if the originating documents are not available. This provides the starting point for applying stylometric classification tasks to plagiarism detection within research articles with limited access to complete document databases.

V. Research Gap

It can be seen from the available literature that the conven-tional approach to plagiarism detection uses string matching algorithms, n-gram overlaps, and fingerprint matching. These conventional methods are quite effective for the copy and paste type of plagiarism. However, the accuracy of the conventional methods decreases when the task involves paraphrased text, formatted text, and translation plagiarism.

Although Natural Language Processing (NLP) techniques and deep learning networks enhanced text analysis, most pre-vious works rely extensively on semantic similarity (content) and less on stylometric analysis (writing style).

Although authorship attribution techniques are present in the literature, most are designed for literary analysis rather than scientific papers. Moreover, most systems are composed of separate modules analyzing text under lexicals or syntax alone and not a combination of both.

There is a lack of literature that provides a comprehensive, automated, and structured approach that encompasses feature extraction in the stylometric analysis of a piece, combined with Machine Learning, in order to detect plagiarism in academic papers that have been significantly revised. The

study fills that gap with the proposition of an intelligent system that uses distinct stylistic patterns in order to detect plagiarism and discrepancies in authorship.

The suggested method emphasizes the unified extraction of multi-level stylometric features, such as lexical, syntactic, and structural traits, within a single automated framework in response to these constraints. The system is designed to handle the formal, domain-specific nature of research writing by utilizing supervised machine learning models that were specifically trained on scientific and academic texts. More accurate identification of subtle authorship inconsistencies resulting from translation, paraphrasing, or AI-assisted content creation is made possible by this integrated methodology. As a result, the suggested system goes beyond current approaches by providing a scalable, flexible, and research-focused solution that can handle complex plagiarism situations in contemporary academic publications.

VI. Proposed System

The proposed system uses a database of research articles to train a machine learning algorithm for detecting plagiarism. The input files are first pre-processed to filter out noise and normalize text. Stylometric features such as lexical character features, sentence-level grammatical structures, and paragraph- level structural features are extracted to create a distinct style signature. The style features are used as input for a supervised learning classification algorithm. If a style is a significant deviation from the style of the alleged writer or if a style belongs to a known plagiarized group, it marks a file as plausible plagiarism

To guarantee accurate detection results, the trained model also assesses the extracted stylometric features using suitable performance metrics like accuracy, precision, recall, and F1-score. Because of the system's ability to analyze documents at various granularities, it can identify particular paragraphs or sections that exhibit stylistic inconsistencies. For source verification, these highlighted areas can then be examined using conventional similarity-based plagiarism tools. The suggested method enhances robustness and reduces false positives by integrating stylometric classification with traditional techniques, making it appropriate for real-world academic settings like research institutions, universities, and journals.

Moreover, the hybrid assessment framework enables informed human judgment through interpretable evidence instead of single similarity scores in isolation. Stylometric cues, such as changes in sentence complexity, part-of-speech distribution, and punctuation, provide contextual explanations of why specific parts are highlighted. This layered analysis, when complemented with traditional plagiarism detection tools focused on source identification, contributes to enhanced reliability of decisions made and reduces the risk of misclassifications brought about by common academic phrasing or domain specific terminology. Therefore, the solution proposed will enable transparent, scalable, and defensible plagiarism evaluation processes that are appropriate for high-stakes academic peer review.

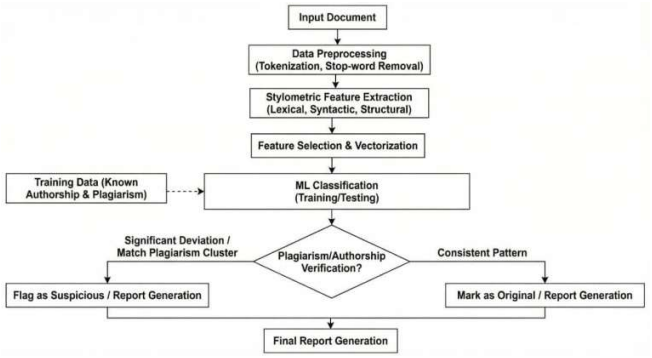

Fig. 1.System architecture of the proposed stylometric plagiarism detection framework

VII. Methodology

The proposed system follows a structured methodology similar to other stylometry-based classification frameworks. It consists of four main stages: data collection, preprocessing, stylometric feature extraction, and machine learning-based classification.

A. System Design

The system is designed as a modular application with the following components:

- Document ingestion layer to upload and store research papers in text format.
- Preprocessing module to segment documents into sections or paragraphs and clean the text.
- Stylometric feature extractor to compute lexical, syntactic, and structural features for each unit of analysis.
- Classification engine that applies trained machine learning models to label sections as "original style" or "suspected plagiarism."
- Reporting interface to visualize style consistency, highlight suspicious regions, and summarize classification results.

All processed feature vectors and prediction outputs are stored in a database to support auditing and further analysis.

B. Data Collection and Preprocessing

The training of this system can be performed using a dataset in which the sections of research papers are labeled as per their authorship or level of plagiarism. The training dataset can consist of the following:

- Works completely written by a solo author referred to as simple documents or genuine documents.
- Mixed documents where some parts are known to be copied, rephrased, or outsourced are classified as mixed.
- Paragraph-level examples with clear markings to distinguish the use of the original author's style from the use of the foreign style.

Preprocessing involves steps such as Tokenization, Sentence Splitting, Stop word removal, and Part-of-speech Tagging. This is done to prepared at a so that meaningful lexical and syntactic variables can be calculated. The process involves segmenting the document into meaningful sections like paragraphs or sections, so that style variation analysis is done on a more detailed level.

C. Stylometric Feature Extraction

The extraction of stylometric features is the vital part of the proposed system. The extracted features are classified into three types:

1) Lexical characteristics:

- Total number of words & distinct words.
- Average length & lexical richness.
- Count of upper-case letters, lower-case letters, numeric characters & special characters.

2) Syntactic:

- Distributions regarding the parts of speech, including nouns, pronouns, adjectives, adverbs, verbs, and determiners.
- Conjunctions, prepositions, modal auxiliaries, and wh-words frequency.
- Occurrence of specific grammatical patterns such as past tense verbs and participles.

3) Structuralaspects:

- Number of sentences per section or paragraph.
- Average number of words per sentence.
- Quotations and Commas

All these features are combined to form a fixed-size feature vector, which denotes the style of the text unit. However, because the features have been measured in different units, scaling normalizations like vector scaling or min-max scaling may be used to normalize all the features to form this combined vector. This enhances the accuracy of the classifiers used in machine learning.

D. Machine Learning Classification

In the classification phase, the work uses both classical and ensemble learning models. Some of these classical models are:

- k-Nearest Neighbors, that assigns a label depending on the majority class among the nearest neighbors in feature space.
- Gaussian Naive Bayes has modeling responsibilities for the probability distribution of features under each class.
- Decision Trees: These learn simple decision rules from stylometric features.

These include:

- XGBoost: This is an efficient, scalable implementation of gradient boosting based on decision trees.
- Stacking: predicts with the meta-classifier using Logistic Regression, k-Nearest Neighbors, Decision Trees, Support Vector Machines, and Naive Bayes as base learners.

It uses labeled feature vectors for training the models, cross- validation for tuning hyper parameters, and to estimate the performance. The most relevant metrics for performance evaluation are accuracy, precision, recall, and F1 score, reflecting how well the system is able to tell apart the genuine from suspicious style.

Once trained, the models can be applied in three modes:

- Document-level mode: classify entire papers as consistent or suspicious.
- It should support a mode at the section level to detect mixed documents that combine several styles.
- Paragraph-level mode to credit authors' original writing and to identify portions of a paper that may contain plagiarism.

1) Step 1: Document Collection and Labeling: Research papers are compiled in the text format and stored as labeled data. The labeled dataset consists of actual papers written by the same author, mixed papers with plagiarized information, or paragraphs where the author has clearly stated their writing.

2) Step 2: Document Ingestion and Storage: Documents uploaded from various sources are ingested in the form of the document ingestion layer and are then stored in a centralized repository. Information such as the ID of the document, the authors of the document, and the section of the document is recorded in order to maintain traceability.

3) Step 3: Text Segmentation: The individual document is segmented into smaller units called sections and paragraphs. These segments enable the detection and investigation of small-scale characteristics and identification of plagiarized and mixed content within a document.

4) Step 4: Text Preprocessing: Preprocessing involves the cleaning of the input data from the natural language input in the required forms according to the standard input formats that can be used to accurately identify the linguistic features from the input.

5) Step 5: Stylometric Feature Extraction: Stylometric characteristics are derived from text units; the characteristics include lexical characteristics, syntactical characteristics, and structural characteristics. They focus on the writing style of an author that is unique to them by highlighting their use of particular words and sentence grammar construction behavior.

6) Step6: Feature Normalization and Vector Construction: Due to differences in the scales of the extracted feature values, normalization techniques like min-max scalar value mapping or vector normalization are performed. All the feature values are then mapped together to a fixed-size feature vector, which represents the stylistic profile of each text unit.

7) Step 7: Author Style Modeling: The actual document samples will then be used to create an author-specific baseline style model, which is a basic concept for comparing changes in an unknown document.

8) Step 8: Model Training: The normalized feature vectors are then fed into a set of various machine learning algorithms such as k-Nearest Neighbors, Naive Bayes, Decision Trees, and Ensemble approaches including XGBoost and stacking classifiers for prediction. Hyperparameter tuning and preven-

tion of overfitting are achieved with cross-validation.

9) Step 9: Classification and Style Consistency Analysis: The trained models will determine if each document, section, or paragraph of the text is original or suspected of plagiarism. Style Consistency Scores are calculated to reveal unexpected changes in style within the text to pinpoint possible foreign or plagiarized content.

10) Step 10: Multi-Level Detection: Three tiers exist in the system:

- Document Level: Detects fully plagiarized papers
- Section Level: Identifies mixed documents that have various writing levels
- Paragraph Level: Identifies localized plagiarized or AI-produced passages

11) Step 11: Aggregation and Reporting: The results are aggregated and visualized through a reporting interface. Suspicious regions are highlighted, and statistics are generated for manual review by the supervisor or reviewer.

12) Step12: Performance Evaluation: System performance can be defined in terms of accuracy, precision, recall, and F1-score. This evaluates the efficiency of the proposed method in distinguishing real writing from plagiarized or foreign writing.

VIII. RESULTS

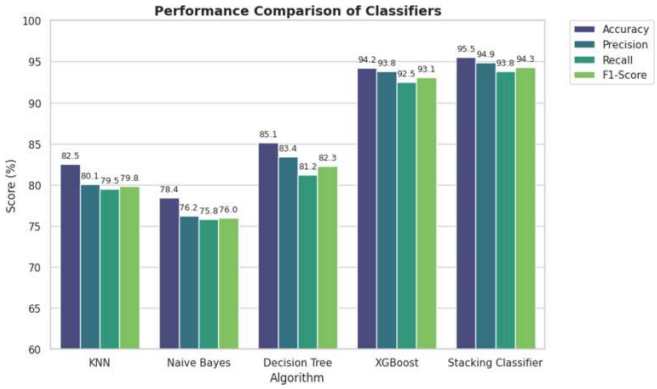

Fig.2.ComparisonofclassificationaccuracyandF1-scoresacrossfivedifferent algorithms

Most of the earlier research that used stylometric feature extraction and machine learning reported high performances in distinguishing writing styles, with performances usually greater than 90 percent. Generally, the ensemble methods like XGBoost and stacking outperform the single classifiers showing strong robustness to noise and variation in text. These findings indicate that a similar approach applied to the study of plagiarism in research papers should be able to achieve reliable detection of style inconsistencies.

It is expected that the stylometric classifiers in the proposed system will:

- Correctly identify documents that are stylistically consistent with the known author profile.
- Detect mixed papers - some sections of these papers deviate significantly from the base style.

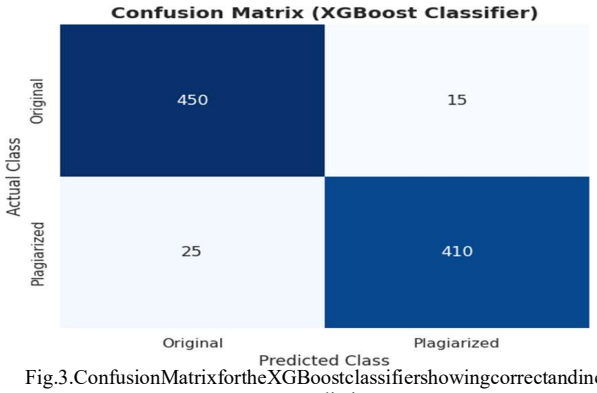

Fig.3.ConfusionMatrixfortheXGBoostclassifiershowingcorrectandincorrect predictions

- Highlight the paragraphs whose stylometric patterns are more similar to external or unknown styles than the claimed author's style.

Such output can be visualized as heat maps or labeled reports showing where the style changes inside a paper. Although this system does not prove plagiarism on its own, it gives strong evidence based on style that guides human reviewers to investigate certain sections with more care.

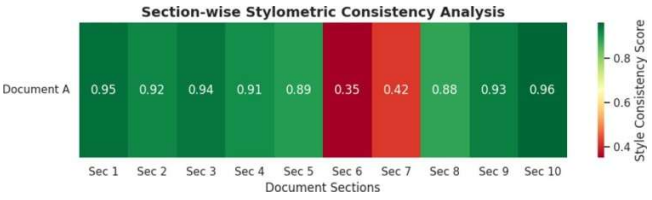

Fig.4.Relativeimportanceofextractedstylometricfeaturesindetectingauthorship style

IX. DISCUSSION

There are a number of benefits using stylometric analysis in plagiarism detection. Firstly, it is not dependent on directly accessing all other potential source materials for a piece but is still able to point out suspicious areas based on stylistic consistency within the piece itself. Secondly, it is not reliant on paraphrases or word-level substitutions but on much deeper stylistic patterns, making it ideal for a new age where content is either pre-computed using software assistance or manually crafted to dodge simple similarity searches.

However, there are also some short comings in the usage of stylometry. Writing style in some cases naturally varies intext, for example, in method and discussion sections, or in collaborative writing by authors for apaper. To overcome

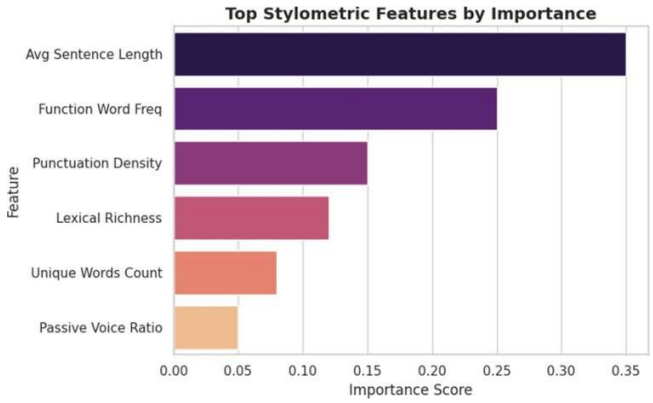

Fig. 5.Section-level stylometric consistency map detecting style deviations in Sections 7 and 8

these, the model has to be trained on proper instances and tested on various documents. If there is any variation in writing style, then the model marks it as a case of style mismatch, but it does not pinpoint the source directly; hence, the model has to be used jointly with the traditional plagiarism detection software.

X. Expected Outcomes

- Precise identification of original and plagiarized documents.
- Computerized extraction of stylometric features such as lexical, syntactic, and structural variables.
- Effective detection of paraphrased and obfuscated text.
- Reduced dependency on manual validation of suspected documents.
- Accurate authorship attribution using Machine Learning classifiers.
- Improved integrity of academic work and better protection of copyright.

XI. Conclusions

This article researches a stylometric feature extraction with machine learning approach to plagiarism detection in scientific articles. Unlike the conventional methods of plagiarism detection whose main objective is to measure direct textual similarity (for example, as it can be evaluated by string matching operation), the proposed system targets authorship style modeling for inconsistency in writing. Stylometry extracts an author's distinctive writing "fingerprint" by using lexical (e.g.,word frequency, vocabulary richness, and n-gram occurrences), syntactic (e.g., sentence length, punctuation pattern, and part-of-speech distribution), and structural features (e.g. paragraph structure, heading usage, or formatting style).

Upon capturing these features, the system can form a complete picture of an author's characteristic writing style. Then, we train a set of machine learning models - including classical classifiers (like Support Vector Machine (SVM), Na¨ıve Bayes, and Logistic Regression) as well as ensemble methods (like Random Forest and Gradient Boosting)-to differentiate between legitimate author-written content and suspicious parts of the text. The approach can be used to detect copying whole document, from paragraph and section level down to the single word level.

This method is especially useful for detecting paraphrased, restructured, or stylistically altered plagiarism, which frequently eludes traditional tools, and it also enhances traditional similarity-based plagiarism detection techniques. Stylometry- based plagiarism detection shows great promise as a depend- able and scalable solution with carefully selected training datasets, well-designed feature sets, and strong evaluation metrics.

Overall, this study shows that plagiarism detection systems can be greatly improved by combining machine learning and stylometric analysis. In a time when more people have access to digital and AI-generated content, this strategy can help universities, academic journals, and educators maintain academic integrity, guarantee originality in scholarly writing, and deter unethical research practices.

References

[1] A. Koubaa, B. Qureshi, A. Ammar, Z. Khan, W. Boulila, L. Ghouti,"Humans are still better than ChatGPT: case of the IEEEXtremecompetition," *Heliyon*, vol. 9, no. 11, Nov. 2023, Art. no. e21624,https://doi.org/10.1016/j.heliyon.2023.e21624.

[2] A.Koubaa,W.Boulila,L.Ghouti,A.Alzahem,S.Latif,"Explor- ingChatGPTCapabilitiesandlimitations:asurvey,"*IEEEAccess*,vol.11,202 3,pp.118698–118721,https://doi.org/10.1109/ACCESS. 2023.3326474.

[3] D.A.Schweidel,M.Reisenbichler,T.Reutterer,K.Zhang,"LeveragingAI FOR CONTENT generation: a customer equity perspective," *Artifi-cial Intelligence in Marketing Review of Marketing Research*, vol. 20,2023, pp. 125–145, https://doi.org/10.1108/S1548643520230000020006.

[4] Y. Feng, P. Poralla, S. Dash, K. Li, V. Desai, M. Qiu, "The impact ofChatGPT on streaming media: a crowdsourced and data-driven analysisusing twitter and reddit," in *2023 IEEE 9th Intl Conference on Big DataSecurity on Cloud (BigDataSecurity), IEEE Intl Conference on HighPerformance and Smart Computing, (HPSC) and IEEE Intl Conferenceon Intelligent Data and Security (IDS)*, May 2023, pp. 222–227, https: //doi.org/10.1109/BIGDATASECURITYHPSCIDS58521.2023.00046.

[5] Z. Jin, Z. Song, "Generating Coherent Comic with Rich Story UsingChatGPTandStableDiffusion,"May2023[Online].Available:https: //arxiv.org/abs/2305.11067v2.(Accessed18June2023).

[6] Z. Jin, Z. Song, "Generating Coherent Comic with Rich Story UsingChatGPTandStableDiffusion,"May2023[Online].Available:https: //arxiv.org/abs/2305.11067v2.(Accessed18June2023).

[7] M.A.Jaber,A.Beganovic´,A.A.Almisreb,"Methodsandapplicationsof ChatGPT in software development: a literature review," *SoutheastEurope Journal of Soft Computing*, vol. 12, no. 1, May 2023, pp. 8–12,https://doi.org/10.21533/SCJOURNAL.V12I1.251.

