IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 September 2023, accepted 20 November 2023, date of publication 1 December 2023,
date of current version 18 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3338855

TOPICAL REVIEW

Exploring the Landscape of Intrinsic Plagiarism Detection: Benchmarks, Techniques, Evolution, and Challenges

MUHAMMAD FARAZ MANZOOR[1], MUHAMMAD SHOAIB FAROOQ[1], MUHAMMAD HASEEB[1], UZMA FAROOQ[1], SOHAIL KHALID[2], AND ADNAN ABID[3], (Senior Member, IEEE)

[1]Department of Computer Science, University of Management and Technology, Lahore 54770, Pakistan
[2]Petroleum Engineering Application Service Department, Saudi Aramco, Dhahran 31311, Saudi Arabia
[3]Department of Data Science, Faculty of Computing and Information Technology, University of the Punjab, Lahore 54590, Pakistan

Corresponding author: Adnan Abid (adnan.abid@pu.edu.pk)

ABSTRACT In the realm of text analysis, intrinsic plagiarism detection plays a crucial role by aiming to identify instances of plagiarized content within a document and determining whether parts of the text originate from the same author. As the development of Large Language Models (LLMs) based content generation tools such as, ChatGPT is publicly available, the challenge of intrinsic plagiarism has become increasingly significant in various domains. Consequently, there is a growing demand for robust and accurate detection methods to address this evolving landscape. This study conducts a comprehensive Systematic Literature Review (SLR), analyzing 44 research papers that explore various facets of intrinsic plagiarism detection, including common datasets, feature extraction techniques, and detection methods. This SLR also highlights the evolution of detection approaches over time and the challenges faced in this context especially challenges associated with low-resource languages. To the best of our knowledge, there is no SLR exclusively based on the intrinsic plagiarism detection that bridge the gap in existing literature and offering valuable insights to researchers and practitioners. By consolidating the state-of-the-art findings, this SLR serves as a foundation for future research, enabling the development of more effective and efficient plagiarism detection solutions to combat the ever-evolving challenges posed by plagiarism in today's digital age.

INDEX TERMS Intrinsic, plagiarism, feature extraction, machine learning, deep learning, evolution, challenges.

I. INTRODUCTION

Intrinsic Plagiarism Detection (IPD) is an important task in text analysis and plagiarism detection. Its goal is to find cases of content copying or unauthorized duplication within single documents [1]. The architecture of IPD is based on the idea of analyzing one document by itself, without having to compare it to other documents or sources [2]. This distinctive approach focuses on scrutinizing the stylistic aspects, writing nuances, and underlying syntactic structures within a text, making it particularly suited for cases where external sources are not accessible or when the copied content is from non-digitized or unavailable sources [2].

The importance of detecting intrinsic plagiarism, particularly in the context of the rapid advancements in publicly available AI content generation tools like ChatGPT, driven by the introduction of Large Language Models (LLMs), cannot be overstated. These sophisticated AI systems have the capability to generate text that closely mimics human writing, making it increasingly challenging to distinguish between genuine content and AI-generated text [3]. This poses a significant risk to the authenticity and originality of written material across various domains, from academic research to journalistic reporting and creative writing. Intrinsic plagiarism detection assumes paramount importance as it focuses on uncovering similarities and discrepancies within individual documents, offering a critical line of defense against the proliferation of AI-generated content that may

The associate editor coordinating the review of this manuscript and approving it for publication was N. Ramesh Babu.

not always adhere to ethical and academic standards [4]. As AI technologies continue to evolve, the role of intrinsic plagiarism detection becomes even more crucial in preserving the integrity and trustworthiness of textual content [5].

Over the years, the domain of IPD has witnessed considerable exploration and research efforts [6]. Scholars have delved into various aspects, including the development of advanced methodologies, the identification of optimal techniques, and the understanding of key challenges. Researchers have identified and discussed various prominent techniques, such as feature extraction methods that encapsulate the linguistic attributes of documents, as well as sophisticated algorithms that distinguish genuine stylistic variations from potential cases of plagiarism [7].

The progress in IPD research is supported by the availability of benchmark datasets tailored to this specific context. These datasets, drawn from diverse sources such as newspapers, articles, reviews, and emails, serve as crucial resources for evaluating and benchmarking IPD techniques [8]. Prior to analysis, pre-processing techniques are employed to ensure data quality and consistency. Techniques encompassing tokenization [9], stop-word removal [10], and stemming [11], are commonly employed to refine the textual content.

Foundation of IPD lies in the extraction of features that encapsulate the document's stylistic attributes [12]. Various techniques like N-grams [13], Vector Space Models (VSM) [14], Latent Semantic Analysis (LSA) [15], and Word Embeddings [16] have been harnessed to distill these features. N-grams, for instance, capture sequential patterns of words or characters, while VSM facilitates the conversion of text into numerical representations, enabling quantitative comparisons. LSA, on the other hand, leverages the semantic relationships between words to unveil hidden patterns [17].

In the pursuit of identifying instances of plagiarism, diverse IPD techniques have emerged, ranging from supervised machine learning to unsupervised methods. These techniques leverage the amalgamation of textual attributes and linguistic patterns to discern instances of potential content replication. The applications of IPD are diverse, encompassing various domains including literature, academic research, and programming code, where safeguarding against unauthorized duplication is paramount.

As IPD techniques continue to evolve, challenges arise, such as the dynamic nature of plagiarism methods and the development of advanced language manipulation technologies [18]. Additionally, the unique challenges encountered in low-resource languages underscore the need for tailored techniques and resources [19]. Moving forward, addressing these challenges while considering the distinctive attributes of low-resource languages holds significant promise for enhancing the accuracy and effectiveness of IPD methodologies.

Notably, while the landscape of plagiarism detection has seen numerous review studies encompassing various aspects such as author identification [2], extrinsic plagiarism detection [20], and other diverse domains of the field, this systematic literature review (SLR) stands out as the pioneering endeavor dedicated exclusively to the realm of Intrinsic Plagiarism Detection (IPD). By focusing solely on IPD techniques, datasets, challenges, and evolution, this review bridges an existing gap in the literature, providing a comprehensive and in-depth analysis that addresses the unique complexities and intricacies of detecting plagiarism solely through inherent features of a single document.

The primary objective of this systematic literature review (SLR) is to comprehensively survey and analyze the landscape of Intrinsic Plagiarism Detection (IPD) techniques. By examining common datasets, exploring various feature extraction methods, tracing the evolution of IPD techniques, and highlighting the major challenges faced in this domain, the aim is to provide a holistic understanding of the advancements, trends, and existing gaps in IPD research. Additionally, this review seeks to shed light on the challenges specific to low-resource languages and to propose potential directions for future research in IPD, ultimately contributing to the enhancement of plagiarism detection strategies and the mitigation of instances of unauthorized content replication. Following are the novelties and contribution of this SLR:

- Comprehensive overview of benchmark datasets commonly used for evaluating Intrinsic Plagiarism Detection (IPD) techniques.
- In-depth analysis of diverse feature extraction techniques employed in IPD, including N-grams, Vector Space Models (VSM), and Word Embeddings.
- Thorough exploration of the evolution of IPD techniques, providing insights into the trajectory of advancements in this field.
- Identification and elaboration of major challenges in IPD, shedding light on the intricacies of distinguishing genuine style variations from instances of plagiarism.
- In-depth examination of the unique challenges encountered in low-resource languages within the context of IPD, emphasizing the scarcity of resources and linguistic diversity.

Rest of the paper is organized as follows: section II highlights the main differences between this literature review and other published literature reviews. Section III discusses the research methodology for this study. Section IV discusses in depth analysis of the papers reviewed in this study. Section V highlights the common evaluation metrics used in intrinsic plagiarism detection tasks. Applications of the intrinsic plagiarism detection are presented in the section VI. Evolution, challenges and way forward are discussed in section VII. Lastly the paper is concluded in section VIII.

II. RELATED WORK

Systematic Literature Reviews (SLRs) have emerged as invaluable method for comprehensively exploring and comprehending various domains. Within diverse fields, a range of insightful review investigations have been conducted [21], [22], illuminating critical aspects. These review articles often fall into distinct categories, including narrative or conventional reviews [23], [24], systematic literature reviews [25], and meta reviews or mapping studies [26], [27]. In the present study, a systematic literature review is presented, delving into the realm of intrinsic plagiarism detection, contributing a unique perspective and analysis to this specific field of study.

A notable absence of a dedicated SLR and review paper exclusively focused on intrinsic plagiarism detection has been observed. This unique gap prompted the undertaking of the present SLR, aiming to comprehensively address this crucial topic. To contextualize the distinctiveness of this study, a comparative analysis with relevant SLRs and survey papers in related domains was conducted.

One key differentiation emerges when examining the scope of the discussed domains. While various existing SLRs and surveys have explored aspects like academic plagiarism [1], [28], authorship verification [29], and writing style change [2], none of these exclusively explored intrinsic plagiarism detection. This absence marks a crucial gap in the literature, which the current study endeavors to fill.

In terms of comprehensiveness, Table 1 provided details the coverage of key aspects across different studies and this SLR. Notably, academic plagiarism and authorship verification have been explored in some existing works. However, aspects like writing style change, common datasets, and challenges associated with intrinsic plagiarism detection were either minimally or not addressed in these prior studies. In contrast, this SLR takes a more comprehensive approach, spanning aspects such as preprocessing techniques, feature extraction techniques, detection techniques, evaluation metrics, evolution, challenges, challenges in low resource languages, and the way forward in the realm of intrinsic plagiarism detection.

Furthermore, the current SLR differs in its in-depth exploration of challenges, encompassing both general challenges and the specific challenges posed by low-resource languages. While some studies like [30] have touched upon challenges, none have explicitly ventured into the intricacies of handling intrinsic plagiarism detection challenges in languages with limited linguistic resources.

Moreover, the inclusion of a "Way Forward" section in this SLR sets it apart. This section envisions potential directions and advancements for the field of intrinsic plagiarism detection, offering valuable insights for future research and development.

Moreover, the inclusion of a "Way Forward" section in this SLR sets it apart. This section envisions potential directions and advancements for the field of intrinsic plagiarism detection, offering valuable insights for future research and development.

The above discussion highlights that the substantial research has been done in various domains of plagiarism detection. These researches have aimed to propose an array of detection methods, develop and evaluate diverse tools, techniques, and strategies, and enhance the field of plagiarism detection. This SLR contributes a comprehensive synthesis of prior research. Table 1 provides a visual comparison between our study and other relevant IPD studies, using checkmarks (✓) and dash symbols (–) to signify "included" and "not included," respectively. In this context, our study embarks on a systematic mapping survey of IPD, with a specific focus on consolidating, categorizing, and extensively discussing the existing body of knowledge concerning IPD techniques, approaches, and other pertinent facets.

TABLE 1. Comparison with other relevant studies.

Aspect\ Reference	[28]	[31]	[1]	[30]	[32]	This Study
Domain	Academic Plagiarism	Authorship Verification	Academic Plagiarism	Academic Plagiarism	Authorship Identification	Intrinsic Plagiarism
SLR	✓	✓	✓	-	-	✓
Common Datasets	-	✓	-	-	-	✓
Preprocessing Techniques	✓	✓	✓	✓	✓	✓
Feature Extraction Techniques	✓	✓	✓	✓	✓	✓
Detection Techniques	✓	✓	✓	✓	✓	✓
Evaluation Metrics	-	-	-	-	✓	✓
Evolution	-	-	-	-	-	✓
Challenges	✓	-	-	-	✓	✓
Limitations in Low Resource Languages	-	-	-	-	-	✓
Way Forward	-	-	-	-	-	✓
Year	2015	2016	2019	2020	2022	2023

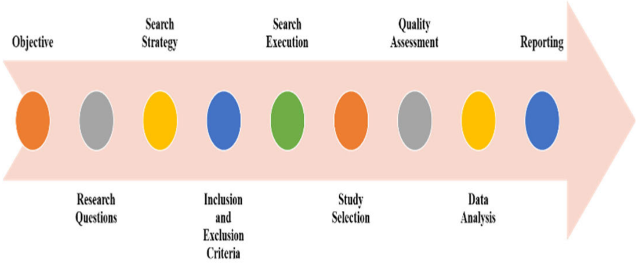

FIGURE 1. Research methodology.

III. RESEARCH METHODOLOGY

To maintain the primary focus of this study, which is to review the research conducted in the area of intrinsic plagiarism detection, we have gathered insights and advice from existing methods described in various studies [21], [33] as shown in Figure 1. By drawing on this knowledge, we have formulated clear research objectives and devised appropriate research questions and search strategies. This approach allows us to effectively search for and identify relevant papers in the field of intrinsic plagiarism detection.

A. RESEARCH OBJECTIVES

The Following are the research objectives of this study:

- To identify and analyze the most common feature extraction techniques employed in intrinsic plagiarism detection.
- To explore and compare the most common techniques utilized in intrinsic plagiarism detection.
- To show the evolution of intrinsic plagiarism detection techniques over time, tracing the development of novel approaches and their impact on improving the accuracy and efficiency of plagiarism detection.
- To identify and examine the challenges faced in intrinsic plagiarism detection.

B. RESEARCH QUESTIONS

To address the research objectives effectively, we have formulated a set of pertinent research questions, each designed to explore specific aspects of intrinsic plagiarism detection. These research questions are presented in Table 2, along with their underlying motivations.

C. SEARCH STRATEGY

The Following search string used to find relevant articles to conduct this study.

(“intrinsic plagiarism”) AND (“detection” OR “analysis”) AND (“feature extraction techniques” OR “feature extraction methods” OR “stylometric features” OR “style analysis” OR “grammar analysis” OR “syntax analysis”) AND (“techniques” OR “methods” OR “approaches” OR “algorithms”) AND (“evolution” OR “development” OR “advancements” OR “progress”) AND (“challenges” OR “limitations” OR “issues” OR “obstacles”).

TABLE 2. Research questions and motivations.

SR#	Research Question	Motivation
RQ1	What are the most prevalent feature extraction techniques used in intrinsic plagiarism detection.	Identifying the most prevalent feature extraction techniques in intrinsic plagiarism detection is essential to understand the current state-of-the-art in the field.
RQ2	How do the commonly employed techniques in intrinsic plagiarism detection compare in terms of their performance?	Comparing the performance of commonly employed techniques in intrinsic plagiarism detection allows us to discern their strengths and weaknesses.
RQ3	How have intrinsic plagiarism detection techniques evolved over time?	Examining the evolution of intrinsic plagiarism detection techniques over time provides insights into the progress and development of the field.
RQ4	What are the major challenges encountered in intrinsic plagiarism detection and how can we move forward to address these challenges effectively?	Identifying and understanding the major challenges in intrinsic plagiarism detection is crucial to move forward in current methods.

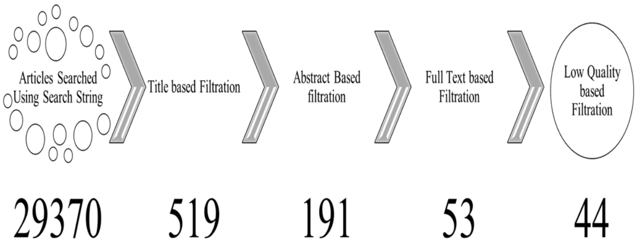
Articles Searched Using Search String
Title based Filtration
Abstract Based filtration
Full Text based Filtration
Low Quality based Filtration
29370
519
191
53
44

FIGURE 2. Study selection process.

The search for primary studies in the field of intrinsic plagiarism detection involves collecting articles from diverse sources, including CLEF, Elsevier, Springer, ArXiv and other reputable journals and conferences.

D. STUDY SELECTION

The study selection is a critical step in the systematic literature review process [34]. It involves reviewing the titles and abstracts of the articles obtained through the search strategy to identify relevant studies that meet the inclusion and exclusion criteria. The goal of this step is to reduce the number of articles to a manageable level while retaining those that are most likely to provide useful information as shown in Figure 2.

TABLE 3. Study selection criteria.

Criteria #	Inclusion criteria	Exclusion Criteria
IE1	Studies focusing on intrinsic plagiarism detection techniques.	Studies not related to intrinsic plagiarism detection.
IE2	Articles published in peer-reviewed journals, conference proceedings, or reputable sources.	Non-peer-reviewed sources or papers from unreliable or questionable publishers.
IE3	Research papers written in English.	Non-English articles
IE4	Studies that provide detailed information on the feature extraction techniques, models, or algorithms used for intrinsic plagiarism detection.	Articles lacking sufficient information on the methodologies or techniques employed.

In this study, a total of 29,370 initial studies were retrieved for intrinsic plagiarism detection from various sources. The selection process involved two authors shortlisting the articles based on predefined inclusion and exclusion criteria. Any conflicts were resolved by involving a third author, and the inclusion/exclusion criteria were refined. The inter-rater agreement between the two authors was found to be almost perfect, with a Cohen's Kappa coefficient of 0.89 [35].

- **Title based search:** In the first step, papers that are unrelated based on their title are carefully eliminated. There were a lot of irrelevant papers at this point. After this step, only 519 papers remained.
- **Abstract based search:** At this step, the abstracts of the articles that were chosen in the earlier stage are examined, and the papers are organised for analysis and research methodology. After this point, there were just 191 papers left.
- **Full text based analysis:** At this stage, the empirical quality of the articles chosen in the earlier stage is assessed. A comprehensive text analysis of the study has been done. From 134 articles, a total of 53 papers were selected.
- **Low Quality Papers:** The final stage of the study selection was to exclude the papers that are not listed in google scholar database. Also, papers that published without Digital Object Identifier(DOI) number, are also excluded from the study. the number of papers in different stages of the selection process for all involved portals has been presented in Table 3.

E. QUALITY ASSESSMENT CRITERIA

Following are the criteria used to assess the quality of the selected primary studies. This quality assessment was conducted by two authors as explained above.

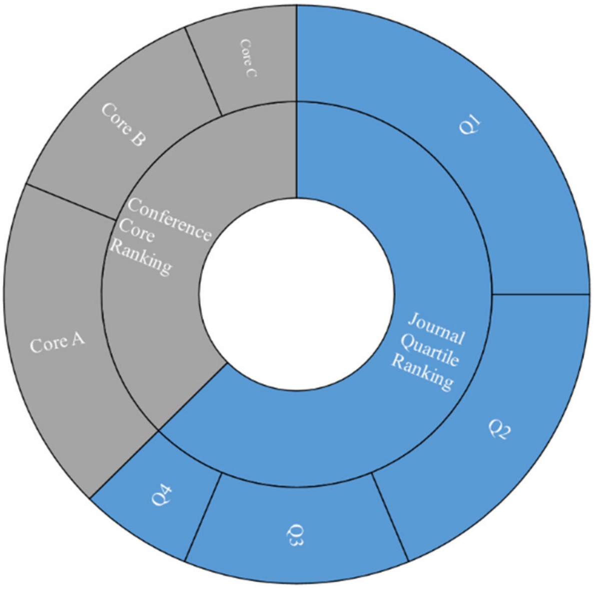

FIGURE 3. Score pattern of publication channels.

a. The study focuses on Integration, Analysis, and Visualization of genomic data, the possible answers were Yes (1), No (0).
b. The study is published recently, the possible answers are, after 2015(2), 2010-2014(1.5), 2005-2010(1) and before 2005(0).
c. The study focuses on empirical results, Yes (1), No (0).
d. The study is published in a well reputed venue that is adjudged through the CORE ranking of conferences, and Scientific Journal Ranking (SJR). The possible answers to this question are given in Figure 3.

F. STUDY SELECTION RESULTS

A total of 44 papers were identified and analyzed for the answers of RQs described above. Table 4 shows the source wise study distribution and Table 5 represents a list of the nominated papers with detail of the classification results and their quality assessment scores.

More 20 papers had scored more than 80% of the score and remaining papers had scored more than mean score i.e., 2.5. Some articles with the score below 2.5 have also been included in this study as they present some useful information and were published in well reputed journals. Also, these studies discuss important demography and technology based aspects that are directly related to intrinsic plagiarism.

IV. ANALYSIS OF COMPILED RESEARCH ARTICLES

In this section, we present the analysis of the selected research articles that have been carefully chosen for this study. The findings are presented in accordance with the research questions outlined in Table 2.

A. DATASETS

This SLR provides a comprehensive exploration of the common datasets utilized for assessing and analyzing the domain

TABLE 4. Study selection results.

Phase	Process	Selection stage	CLEF	Elsevier	Springer	IEEE	SemEval	PAN	Others	Total
1	Search	Search string	250	400	2130	4535	3597	2313	6975	20200
2	Screening	Title	69	39	74	67	99	57	114	519
3	Screening	Abstract	43	26	27	17	30	20	28	191
4	Screening	Full text	9	3	11	2	3	5	20	53
5	Finalizing	Low Quality	9	3	11	2	3	5	11	44

of intrinsic plagiarism detection as listed in Table 6. These datasets play a pivotal role in the realm of intrinsic plagiarism detection, facilitating the evaluation and benchmarking of various detection techniques. These datasets encompass a diverse range of textual content, representing different genres and domains. Notably, the PAN (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection) series of shared tasks have been instrumental in advancing the field [39], [47], [58], [64], [72]. The PAN datasets, spanning multiple years, encompass intrinsic plagiarism detection scenarios, serving as valuable resources for researchers. Another significant dataset is the CEN (Corpus of English Novels) [49], offering a unique collection for evaluating intrinsic plagiarism detection techniques in the context of literature. These datasets, along with others like Wikipedia and diverse web documents, collectively contribute to enhancing the robustness and effectiveness of intrinsic plagiarism detection methods. Their availability empowers researchers to develop, compare, and fine-tune algorithms that can tackle real-world instances of intrinsic plagiarism across various textual sources and styles.

B. PRE PROCESSING AND DATA CLEANING

Intrinsic plagiarism detection approaches require careful preprocessing operations to remove noise while retaining essential information for analysis [56]. However, it is advisable to limit preprocessing steps to a minimum to avoid losing potentially useful information.

Common preprocessing steps for text-based intrinsic plagiarism detection include lowercasing, punctuation removal, tokenization, segmentation, number removal or replacement, named entity recognition, stop words removal, stemming or lemmatization, Part of Speech (PoS) tagging, and synset extension [76]. These steps aim to standardize the text and extract relevant linguistic features for plagiarism analysis. For instance, lowercasing is applied to convert all characters to lowercase to ensure consistency in the text representation [77]. Punctuation removal eliminates punctuation marks from the text, although intrinsic detection methods often retain punctuation to preserve potential indicators of similarity [78].

Tokenization involves splitting the text into individual tokens, such as words or subwords, for further analysis [79]. Segmentation focuses on identifying boundaries between sentences or other linguistic units within the document [79]. Number removal or replacement replaces numerical values with placeholders to reduce the impact of specific numerical content on plagiarism analysis. Also, Named Entity Recognition(NER) is employed to identify and classify named entities, such as people, organizations, or locations, which can provide valuable information for plagiarism detection [17]. Removing stop words involves eliminating common words (e.g., "the," "is," etc.) that do not carry significant semantic meaning [29].

On the other hand, Stemming or lemmatization reduces words to their base form (stem or lemma) to unify related word forms and facilitate comparisons [79]. PoS tagging assigns grammatical categories (e.g., noun, verb, adjective) to each word, aiding in syntactic analysis and identifying potential plagiarism through similar word usage [80]. Additionally, synset extension involves expanding the vocabulary by adding synonyms and related words based on lexical databases or word embeddings, enhancing the coverage of potential similarities [81].

Standard NLP software libraries such as the Natural Language Toolkit (NLTK) in Python or the Stanford CoreNLP library in Java provide well-established tools to perform these preprocessing steps in a convenient and reliable manner. Researchers predominantly utilize these libraries in their intrinsic plagiarism detection studies due to their multilingual and multifunctional text processing pipelines. By carefully selecting and applying these preprocessing steps, intrinsic plagiarism detection methods can effectively clean the text while retaining essential linguistic features needed for accurate analysis.

C. FEATURE EXTRACTION TECHNIQUES

Feature extraction plays a crucial role in intrinsic plagiarism detection. Various techniques and methods are employed to transform the textual content into numerical representations that capture relevant information. A comprehensive taxonomy of commonly used feature extraction techniques are illustrated in Figure 4. These features are then used to compare and analyze the text for similarities, discrepancies, and potential instances of plagiarism. A comprehensive taxonomy of commonly used feature extraction techniques are illustrated in Figure 4.

1) LEXICAL

Lexical-based feature extraction is a common approach used in plagiarism detection, particularly in the context of intrinsic

TABLE 5. Meta detail of selected primary study.

Ref	Year	Channel	Dataset Language	Quality Assessment				
				a	b	c	d	Total
Supervised Machine Learning and Ensemble Techniques								
[36]	2020	CLEF (Working Notes)	English	1	1	1	1	4
[7]	2021	CLEF (Working Notes)	English	1	1	1	1	4
[37]	2019	International Journal of Advanced Computer Science and Applications	English	1	1	1	1	4
[38]	2019	Artificial Intelligence: Methodology, Systems, and Applications	English	1	1	1	1	4
[39]	2014	EMNLP	English, Spanish, Arabic	1	0.5	1	1.5	4
[40]	2015	pecial Interest Group on Discourse and Dialogue	English	1	1	1	0	3
[41]	2017	Engineering Applications of Neural Networks	English	1	1	1	1	4
[42]	2010	Irish Conference on Artificial Intelligence and Cognitive Science	English	1	1	0	0	2
[43]	2019	Future Generation Computer Systems	English	1	1	1	2	5
[44]	2021	CLEF (Working Notes)	English	1	1	1	1	4
[8]	2016	CLEF (Working Notes)	English	1	1	1	1	4
[45]	2013	Chinese Lexical Semantics	English	1	0.5	1	0	2.5
[46]	2010	Language Resources and Evaluation	English	1	1	1	2	5
[47]	2017	Future Generation Computer Systems	English	1	1	1	2	5
[48]	2017	arXiv	English, French	1	1	1	0	3
[49]	2016	SemEval@ NAACL-HLT 2016	English, Spanish	1	0.5	1	1	3.5
[50]	2016	SemEval-2016	English	1	0.5	1	1	3.5
[51]	2018	Evolving systems	English	1	1	1	1.5	4.5
[52]	2013	Notebook for PAN at CLEF Working Notes	English, Greek, Spanish	1	1	1	1	4

TABLE 5. *(Continued.)* Meta detail of selected primary study.

Unsupervised Machine Learning								
[53]	2019	CLEF (Working Notes)	English	1	1	1	1	4
[54]	2017	CLEF (Working Notes)	English	1	1	1	1	4
[55]	2021	ICCCIS	English	1	0.5	1	0	2.5
[56]	2012	CLEF (Online Working Notes/Labs/Workshop)	English	1	0.5	1	1	3.5
Statistical and Distance-based Approaches								
[57]	2011	Proceedings of the PAN	English	1	0.5	1	0	2.5
[58]	2014	ICAISC	English	1	1	1	1	4
[59]	2013	Expert Systems with Applications	English	1	1	1	2	5
[60]	2009	threshold	English	1	1	1	0	3
[61]	2011	Knowledge-Based and Intelligent Information and Engineering Systems	English	1	1	1	1	4
[62]	2012	LNCS	English	1	1	1	0	3
[63]	2012	Natural Language Processing and Information Systems	English	1	1	1	1	4
[64]	2011	Proceedings of the PAN	English	1	1	1	0	3
[65]	2006	Advances in Information Retrieval	English	1	0.5	1	1.5	4
[12]	2018	CLEF (Working Notes)	English	1	1	1	1	4
[66]	2015	Work. Notes PAN-AraPlagDet FIRE 2015	Arabic	1	0.5	1	0	2.5
Classical and Traditional Techniques								
[67]	2017	CLEF (Working Notes)	English	1	1	1	1	4
[14]	2011	Notebook for PAN at CLEF	English	1	0.5	1	1	3.5
[68]	2017	CLEF (Working Notes)	English, Dutch, and Greek	1	1	1	1	4
[69]	2010	Notebook papers of CLEF 2010 LABs and workshops	English	1	1	1	1	4
[70]	2012	Spanish Conference on Information Retrieval	Arabic	1	0.5	1	0	2.5

TABLE 5. *(Continued.)* **Meta detail of selected primary study.**

[71]	2007	Advances in Data Analysis	English	1	1	1	0	3
[72]	2013	Conference on Developments in eSystems Engineering	English	1	1	1	0	3
[73]	2013	Natural Language Processing and Information Systems	English	1	1	1	1	4
[74]	2013	Datenbanksysteme für Business, Technologie und Web	English	1	1	1	0	3
[75]	2016	SemEval-2016	English	1	0.5	1	1	3.5

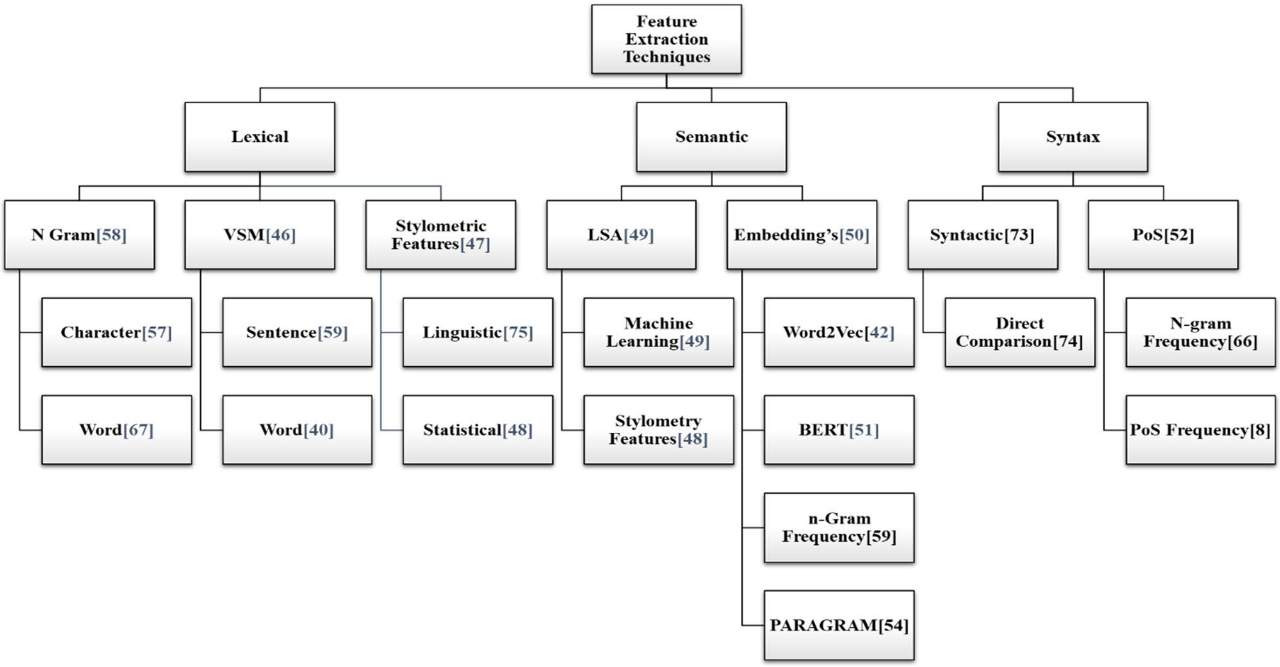

FIGURE 4. **Taxonomy of feature extraction techniques mentioned in the literature.**

plagiarism detection where the focus is on analyzing the text within a single document. Lexical features capture information related to the vocabulary, word usage, and patterns in the text, which can help identify similarities and discrepancies indicative of plagiarism. In lexical-based plagiarism detection, different approaches can be categorized into three main categories: n-gram comparisons, vector space models and stylometric features.

- **N gram**

Researchers in the field of intrinsic plagiarism detection have extensively employed N-grams, which are sequences of N words or characters extracted from a document [58]. The choice of N, such as unigrams, bigrams, or trigrams, is determined based on the desired level of granularity for capturing patterns and similarities within the text [57]. Character-level N-grams, representing sequences of N characters, allow for more profound analysis of the text's structure [39]. Conversely, word-level N-grams encapsulate N-word sequences and offer insights into the document's semantic and syntactic patterns.

Once the N-grams are extracted, they require transformation into a suitable format for analysis. Typically, numerical representations are applied to N-grams for quantitative comparisons and calculations. Techniques like one-hot encoding or term frequency-inverse document frequency (TF-IDF) are

TABLE 6. Summary of common datasets used in IPD.

Dataset Name	Size	Language	Used in	Source
Benchmark Datasets				
PAN 2009	3092 documents	English	[39], [47],[58], [64], [72]	https://pan.webis.de/sepln09/pan09-web/intrinsic-plagiarism-detection.html
PAN 2010	11148 documents	English	[61], [14]	https://pan.webis.de/clef10/pan10-web/plagiarism-detection.html
PAN 2011	4753 English documents	English	[74],[73], [37],[41],[63],[67] , [57], [58],[14], [8] [64], [72]	https://pan.webis.de/clef11/pan11-web/intrinsic-plagiarism-detection.html
PAN 2012	71 Documents	English	[56],[62]	https://pan.webis.de/clef12/pan12-web/
PAN 2013	10,20,5 Documents	English, Greek, Spanish	[52]	https://pan.webis.de/clef13/pan13-web/
PAN 2016	71 Documents	English	[8], [72]	https://pan.webis.de/clef16/pan16-web/
PAN 2017	187 Documents	English	[68],[75], [59]	https://pan.webis.de/clef17/pan17-web/index.html
PAN 2018	4472 Documents	English	[37], [12], [38]	https://pan.webis.de/clef18/pan18-web/index.html
PAN 2019	3818 Documents	English	[53]	https://pan.webis.de/clef19/pan19-web/index.html
PAN 2020	17380 Documents	English	[36]	https://zenodo.org/record/3660984
PAN 2021	13600 Documents	English	[7],[44]	https://pan.webis.de/clef21/pan21-web/index.html
Other Datasets				
MED	1033 documents	English	[48]	
Corpus of English Novels	292 Novel	English, Spanish	[49]	https://perswww.kuleuven.be/~u0044428/cen.htm
English-Spanish parallel corpus	503474 sentences	English, Spanish	[51]	https://aclanthology.org/2005.mtsummit-papers.11.pdf
STS2016	15618 Sentences	English	[50]	https://paperswithcode.com/dataset/semantic-textual-similarity-2012-2016
IPAT-DC and IPAT-CC	6182 documents each	English	[60]	Private
Manually Created Corpus	450 documents	English	[65]	Private
InAra	1024 Documents	Arabic	[39]	https://sourceforge.net/projects/inaracorpus/
Arabic intrinsic plagiarism detection	10 Documents	Arabic	[70]	Private

utilized, resulting in vector representations of N-grams with binary values or weighted frequencies [8]. Following the representation stage, the subsequent step involves comparing and analyzing the N-grams for similarities and discrepancies. Depending on specific data requirements, cosine similarity, Jaccard similarity, or edit distance are employed as similarity

TABLE 7. Example of N gram technique.

Original Sentence(bi-gram)	*Modified Sentence(bi-gram)*
"The quick"	*"A speedy"*
"quick brown"	*"speedy red"*
"brown fox"	*"red fox"*
"fox jumps"	*"fox leaps"*
"jumps over"	*"leaps over"*
"over the"	*"over the"*
"the lazy"	*"the lazy"*
"lazy dog"	*"lazy dog"*

TABLE 8. Lexical based feature extraction methods.

Method Variation	**Method Extension**	**Papers**
N Gram	Character N Gram	[58], [57], [39], [67]
	Word N Gram	[8], [45]
Vector Space Model	Sentence	[46], [59]
	Word	[40]
Stylometric Features	semantic features	[72]
	Syntactic, PoS, Text statistics	[47], [75]
	LSA	[48]

measures to quantify overlap or similarity between distinct N-gram sets [45].

The incorporation of N-gram-based feature extraction holds considerable significance within the realm of intrinsic plagiarism detection. By capturing the sequential relationships between words within sentences, N-grams provide a valuable tool for identifying subtle similarities or deviations in writing style. This technique enables the identification of patterns and linguistic nuances that may serve as indicators of potential plagiarism instances, even when the wording has been modified.

Following example (Table 7) illustrates how N-gram works in terms of intrinsic plagiarism detection tasks:

- **Original Sentence:** "The quick brown fox jumps over the lazy dog."
- **Modified Sentence:** "A speedy red fox leaps over the lazy dog."

By comparing the N-gram features of these two sentences, we can identify that they share similar bi-grams like "fox leaps" and "over the," indicating potential similarity in writing style.

- **Vector Space Model(VSM)**

Various researchers have extensively utilized the Vector Space Model (VSM) as a fundamental technique in the field of intrinsic plagiarism detection. The VSM method serves to represent documents and assess their similarity through the transformation of textual data into numerical vectors. This approach facilitates quantitative analysis, making it feasible to identify potential instances of plagiarism. The concept behind the VSM involves representing each document as a vector in a multi-dimensional space, with each dimension corresponding to a distinct word or term. The value of each dimension signifies the significance or frequency of the term within the document itself [46].

For assigning weights to terms within the document vectors, the widely employed Term Frequency-Inverse Document Frequency (TF-IDF) technique is employed. Term Frequency (TF) gauges the term's frequency within a specific document, while Inverse Document Frequency (IDF) captures the rarity of the term across the entire corpus of documents. By using TF-IDF, the technique highlights crucial terms within a document while minimizing the impact of frequently occurring terms.

TABLE 9. Example of VSM technique.

Original Sentence	*Modified Sentence*
"The": 1	*"Rainforests": 1*
"climate": 1	*"have": 1*
"of": 1	*"a": 1*
"the": 1	*"warm": 1*
"rainforest": 1	*"and": 1*
"is": 1	*"humid": 1*
"humid": 1	*"climate": 1*
"and": 1	*"Rainforests": 1*
"warm": 1	

Subsequently, after representing documents as vectors, similarity between these vectors is assessed using metrics like cosine similarity. This measure calculates the cosine of the angle between two vectors, producing a value within the range of 0 to 1. A value of 1 signifies nearly identical or highly similar vectors. By gauging the similarity among document vectors, it becomes possible to detect instances of potential textual overlap or plagiarism within a single document [59].

The utilization of the vector space model for intrinsic plagiarism detection allows for quantitative analysis of textual content and the identification of resemblances among documents. This method capitalizes on TF-IDF weighting and vector representation, thereby encapsulating the comprehensive content and distribution of words. This systematic approach proves beneficial in detecting potential instances of plagiarism embedded within a single document [40].

Following is an example(Table 9) of Vector Space Model (VSM) based feature extraction in the context of intrinsic plagiarism detection:

- **Original Sentence**: "The climate of the rainforest is humid and warm."
- **Modified Sentence**: "Rainforests have a warm and humid climate."

Comparing the VSM feature vectors of these two sentences, we can observe that they have similar patterns of term frequencies, indicating potential stylistic similarity.

- **Stylometric Features**

Researchers have widely employed stylometric-based feature extraction techniques in the domain of intrinsic plagiarism detection to delve into the intricate nuances of writing styles within textual content. This approach aims to capture distinctive writing patterns, linguistic nuances, and individualized authorship styles that play a pivotal role in distinguishing authors or detecting instances of text replication.

Stylometric features encapsulate a range of writing style elements, including vocabulary richness, sentence lengths, punctuation utilization, grammatical structures, and syntactic arrangements. These features are meticulously quantified to establish a stylometric profile unique to each document [75]. One prevalent avenue in stylometric feature extraction involves the computation of statistical measures like word frequencies, average sentence length, or the distribution of punctuation marks [47]. These measures offer insights into the author's writing style and inclinations.

Another essential facet of stylometric-based feature extraction revolves around the scrutiny of function words—compact words serving grammatical or syntactical roles rather than conveying explicit meanings. Analyzing the frequencies and patterns of function words, such as articles, pronouns, or prepositions, can serve as a reliable indicator of an author's writing style or textual resemblances [47].

The extraction and analysis of stylometric features enable intrinsic plagiarism detection methodologies to unveil subtle writing patterns and linguistic clues that may suggest instances of plagiarism or text duplication. By harnessing stylometric-based feature extraction, these methods enhance their ability to uncover the distinct writing styles of authors and facilitate the identification of potential textual commonalities or disparities [48]. Here's an example of Stylometric Features based feature extraction using a sentence(Table 10):

- **Sentence:** "The quick brown fox jumps over the lazy dog."

TABLE 10. Example of stylometric feature extraction.

Type	*Features and statistics*
Word Frequency:	*The(2), quick(1), brown(1), fox(1), over(1), lazy(1), dog(1)*
Sentence Length:	*9*
Average Word Length:	*3.89*
Punctuation Usage:	*periods, commas, and exclamation marks etc*
Part of Speech Distribution:	*Noun, verbs, adjectives etc*

2) SEMANTIC

Semantics-based methods in intrinsic plagiarism detection operate on the premise that the similarity between two passages is influenced by the presence of similar semantic units within them [82]. The notion of semantic similarity is derived from the observation that units sharing similar contexts tend to exhibit higher semantic similarity. To leverage semantics in the analysis, many methods utilize thesauri such as WordNet or EuroVoc [81]. These thesauri provide valuable semantic features such as synonyms, hypernyms (superordinate terms), and hyponyms (subordinate terms), which enhance the performance of paraphrase identification. Incorporating these semantic features enables the detection of synonym replacement obfuscation and reduces the dimensionality of the vector space representation. Proper sentence segmentation and text tokenization are crucial steps for semantics-based detection methods. Sentence segmentation ensures that passages are appropriately divided into sentences, while text tokenization extracts the atomic units of analysis, which are typically words or phrases. In most research papers, the prevalent choice for tokens is individual words. The semantic based feature extraction techniques are divided in two categories as shown in Table 11.

TABLE 11. Semantic based feature extraction methods.

Method Variation	**Method Extension**	**Papers**
Latent Semantic Analysis(LSA)	LSA with Machine Learning	[49]
	LSA with Stylometric Features	[48]
Word Embedding's	word2vec	[50], [42]
	Compositional Bilingual Word Embeddings	[51]
	Embeddings with n-gram frequency	[59]
	PARAGRAM-SL999, PARAGRAM-PHRASE	[54]

- **Latent Semantic Analysis(LSA)**

Researchers in the field of intrinsic plagiarism detection have harnessed the power of Latent Semantic Analysis (LSA), a feature extraction technique designed to delve into the inherent semantic meaning embedded within textual content. LSA operates by leveraging the statistical properties of word co-occurrence patterns to construct a semantic representation of documents, effectively capturing their underlying semantic essence.

To gauge the similarity of term distributions across texts, LSA initiates by creating a term-document matrix. This matrix uniquely allocates rows to individual words and columns to documents, subsequently populating it with relevant metrics such as term frequencies or TF-IDF scores [49]. The process then advances to singular value decomposition (SVD), a mathematical technique that dissects the term-document matrix into three distinctive matrices: a term matrix, a singular value matrix, and a document matrix [83]. This decomposition serves the dual purpose of dimensionality reduction and preservation of the most significant latent semantic information.

The outcome of this dimensionality reduction, as facilitated by SVD, captures the intricate semantic relationships interwoven within words and documents. As words closely linked in meaning frequently co-occur in comparable contexts, LSA aptly projects the original documents onto a reduced-dimensional semantic space. This endeavor culminates in the creation of feature vectors that eloquently encapsulate the semantic essence inherent to each document [48].

For example, consider two sentences:

- **Sentence 1**: "The cat chased the mouse."
- **Sentence 2:** "The feline pursued the rodent."

Both sentences convey similar meanings but use different words. LSA-based feature extraction can represent these sentences in a lower-dimensional semantic space, where they are more comparable. By capturing the underlying semantic structure, LSA allows the algorithm to recognize similarities between sentences beyond the exact word matches. This aids in identifying potential instances of plagiarism, even when authors rephrase sentences using different words but retain the same underlying meaning.

- **Word Embedding's**

Word embeddings-based feature extraction is a widely embraced method in intrinsic plagiarism detection, designed to capture the intricate semantic and contextual essence of words within a document. This technique functions by representing words as dense vectors within a continuous vector space, thoughtfully positioning similar words closer together. This proximity-based representation allows for the extraction of significant features that mirror the semantic relationships nestled within words.

The process commences with the training of a word embedding model on an extensive text corpus. During this training phase, the model acquires the art of encoding the semantic meaning of words by diligently analyzing the contexts in which these words appear [50]. Prominent word embedding algorithms, such as Word2Vec or GloVe, rely on techniques like skip-gram or co-occurrence matrix factorization to craft these embeddings [50].

Following the training of the word embedding model, the embeddings for the words in a document are readily obtained by referencing their respective vectors in the embedding space. These vectors, often densely packed with information, encapsulate the word's semantic nuances and contextual relevance [59]. By amalgamating the embeddings of all words within a document, a feature vector materializes, mirroring the document's semantic substance [54]. This approach, fortified by the wealth of semantic and contextual information distilled in word embeddings, amplifies the capacity to unearth instances of plagiarism characterized by semantic manipulation or rewriting. It furnishes a richer representation of the document's essence, thereby facilitating a more profound analysis of semantic congruences and deviations amongst documents.

TABLE 12. Syntax based feature extraction methods.

Method Variation	Method Extension	Papers
Syntactic	Direct Comparison	[73],[74]
PoS Tagging	PoS n-gram frequency	[52],[66]
	PoS frequency	[8]

The following two sentence illustrates the mechanism of word embedding's:

- **Sentence 1:** "The weather is sunny today."
- **Sentence 2:** "Today's weather is sunny."

Word embeddings can map these sentences into high-dimensional vectors. Despite the variations in word order, the embedded vectors for "weather" and "sunny" would likely exhibit similarity, as they convey similar concepts. This similarity allows the detection algorithm to identify potential plagiarism instances, even when authors rephrase sentences while maintaining the same underlying ideas.

3) SYNTAX

Syntax-based feature extraction is an approach commonly used in intrinsic plagiarism detection to analyze the structural patterns and grammatical characteristics of text. Unlike lexical-based methods that focus on word usage and vocabulary, syntax-based approaches delve into the syntactic structure and arrangement of sentences within a document. These features can provide valuable insights into the organization and composition of the text, aiding in the identification of potential instances of plagiarism. Syntactic and PoS are the two categories of the syntax based feature extraction techniques as shown in Table 12.

- **Syntactic**

Syntactic-based feature extraction stands as a pivotal facet within the realm of intrinsic plagiarism detection, dedicated to uncovering the structural blueprints and grammatical makeup of text. This approach transcends mere lexical considerations, diving into the intricate syntax that governs sentence arrangement within a document [84].

Central to this method is the process of sentence parsing, a linguistic endeavor entailing the dissection of sentence grammatical structures and the discernment of word and phrase relationships. This parsing procedure delves deep into syntactic patterns and dependencies, enabling the capture of text organization and composition in a more incisive manner [73].

Furthermore, a host of syntactic features, such as dependency relations and phrase structures, are harnessed. Dependency relations, for instance, capture the syntactic interplay between words, encompassing subject-verb-object connections and proffering insights into sentence structure and similarity. Analysis of phrase structures, including noun phrases and verb phrases, contributes an understanding of syntactic alignment within a sentence. These features collectively facilitate the comparison and identification of akin or paraphrased sentences, lending potency to the detection of instances of plagiarism hinged on syntactic manipulations [74].

For instance:

- **Original Sentence:** "The dog chased the ball."
- **Plagiarized Sentence:** "The ball was chased by the dog."

The syntactic structure captures the relationship between "dog," "chased," "ball," and "was chased by." Analyzing such structures can reveal differences in how authors manipulate sentence syntax while maintaining underlying meaning, aiding in plagiarism identification.

- **PoS**

The utilization of Part-of-Speech (PoS) tagging, a foundational task in the realm of natural language processing (NLP), entails assigning grammatical labels to words within a sentence [52]. These tags offer insight into the syntactic category or role of each word, shedding light on its function in the sentence structure [66].

Intrinsic plagiarism detection benefits significantly from PoS tagging, which assumes a pivotal role in uncovering potential instances of plagiarism by dissecting syntactic structures and disparities between documents. Through the comparison of PoS tags and syntactic configurations, nuances in sentence organization are exposed, potentially revealing textual parallels or discrepancies [8].

- **Original Sentence:** "The cat jumped over the fence."
- **Plagiarized Sentence:** "A fence was jumped over by the cat."

By analyzing the PoS tags of words ("The" as determiner, "cat" as noun, "jumped" as verb, etc.), we can capture the grammatical structure. Detecting changes in PoS patterns, like switching from active to passive voice, aids in uncovering potential plagiarism attempts while retaining the sentence's overall meaning.

D. TECHNIQUES

Intrinsic plagiarism detection techniques encompass a diverse set of approaches that aim to identify plagiarized sections within text documents without relying on a reference corpus as shown in Figure 5. These techniques can be broadly categorized into different groups based on their underlying methodologies. Supervised machine learning and ensemble learning methods involve training classifiers on labeled data to distinguish between plagiarized and original text segments. Unsupervised machine learning techniques, on the other hand, use clustering or outlier detection algorithms to identify suspicious segments without the need for labeled data. Statistical and distance-based approaches rely on quantifying stylistic variations and similarities within a document to detect potential instances of plagiarism. Additionally, other techniques, such as text mining, deep latent semantic analysis, and language-independent stylometric features, contribute to the advancement of intrinsic plagiarism detection methods by leveraging various linguistic and statistical properties of the text.

1) SUPERVISED MACHINE LEARNING TECHNIQUES

Intrinsic plagiarism detection utilizes various supervised machine learning techniques to analyze and identify instances of plagiarism within a single document. These techniques include lexical analysis with n-gram frequencies, embedding-based approaches using word embeddings, and stylometric analysis that focuses on stylistic features. Machine Learning techniques enable the detection of changes in writing style and the identification of potentially plagiarized passages within a document. Moreover, these ML techniques play a crucial role in enhancing the accuracy and effectiveness of intrinsic plagiarism detection methods, providing valuable insights into the detection and prevention of plagiarism in textual content.

Authors have utilized various machine learning techniques that takes lexical based features as input for the intrinsic plagiarism detection. Such as, Kopev et al. [38] presents a supervised approach for style change detection, which involves predicting whether style changes occur in a text document and identifying the specific positions of these changes. The authors combine a TF.IDF representation of the document with task-specific engineered features. Predictions are made using an ensemble of diverse classifiers, including SVM, Random Forest, AdaBoost, MLP, and LightGBM. Recursive application of the model is performed to locate the exact positions of style changes. The proposed approach achieved the winning performance in the PAN@CLEF 2018 task on Style Change Detection. Moreover, Bensalem et al. [39] presents a novel language-independent intrinsic plagiarism detection method based on a new text representation called n-gram classes.

The proposed method is evaluated on three publicly available standard corpora and achieves comparable results to the best state-of-the-art methods. The authors introduce the concept of n-gram classes as a way of quantifying the writing style. The method uses a supervised classification-based approach and demonstrates the ability to discriminate between plagiarized and original text fragments, despite using a small number of features when building the classification model. Also, Rahman [67] presents information theoretical and statistical features, including function word skip n-grams, for intrinsic plagiarism detection. A binary classifier is trained using different feature sets, and a set

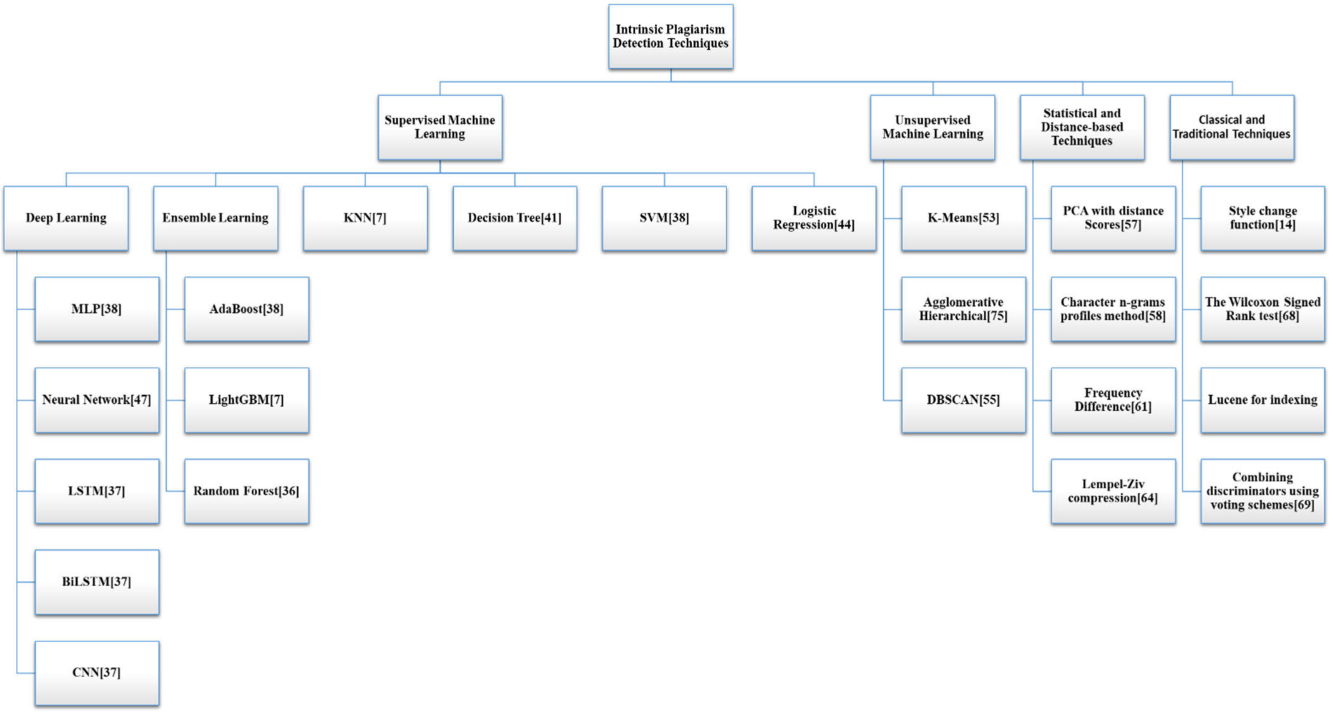
Intrinsic Plagiarism Detection Techniques
Supervised Machine Learning
Unsupervised Machine Learning
Statistical and Distance-based Techniques
Classical and Traditional Techniques
Deep Learning
Ensemble Learning
KNN[7]
Decision Tree[41]
SVM[38]
Logistic Regression[44]
K-Means[53]
PCA with distance Scores[57]
Style change function[14]
MLP[38]
AdaBoost[38]
Agglomerative Hierarchical[75]
Character n-grams profiles method[58]
The Wilcoxon Signed Rank test[68]
Neural Network[47]
LightGBM[7]
DBSCAN[55]
Frequency Difference[61]
Lucene for indexing
LSTM[37]
Random Forest[36]
Lempel-Ziv compression[64]
Combining discriminators using voting schemes[69]
BiLSTM[37]
CNN[37]

FIGURE 5. Taxonomy of IPD techniques mentioned in the literature.

of 36 features is proposed for classifying plagiarized and non-plagiarized text segments.

On the other hand, authors have also used stylometric features as an input in ML based approaches. For instance, Polydouri et al. [41] introduces a new approach that incorporates machine learning and considers the challenge of unbalanced training datasets in this context. The proposed detection system is evaluated using the PAN Webis intrinsic plagiarism detection corpora from 2009 and 2011. Moreover, Curran [47] presents an evolutionary neural network approach for intrinsic plagiarism detection, showcasing its effectiveness in developing classifiers. The approach does not require reference collections and is applied to identify plagiarized sections within a document. Also, AlSallal et al. [43] presents a novel approach for intrinsic plagiarism detection without reference collections. It combines statistical properties of common words, latent semantic analysis, stylometry, and MLP neural networks. The method focuses on authorial attributes and utilizes the Corpus of English Novels dataset for evaluation. Furthermore, Singh et al. [44] presents a machine learning approach for the PAN 2021 Style Change Detection Task. The task is transformed into an authorship verification problem, where stylometric features are extracted from paragraphs and used to train a Logistic Regression classifier. The model is then applied to predict style changes in three different style change detection tasks. The approach is based on a modified version of the authorship verification method used in a previous PAN competition.

Different types of embeddings have also been used in the context of intrinsic plagiarism detection to capture the semantic and syntactic information of text documents. BERT embeddings, a popular language representation model, have been employed to generate contextualized word embeddings that capture the meaning of words in the context of the entire sentence. Constituency Trees Embedding (CTE) and Dependency Trees Embedding (DTE) are other techniques used to convert sentences into structured representations, where CTE represents sentences as constituency trees and DTE represents them as dependency trees. Iyer and Vosoughi [36] proposed a method for Style Change Detection that utilizes BERT, a pretrained bidirectional language model by Google AI, for tokenization and generating embeddings of sentences. These embeddings are then used to train a random forest classifier. Also, Strøm [7] presents a solution to the PAN 2021 shared task on style change detection, which involves multi-label multi-output classification. The authors propose a pragmatic approach, utilizing binary classification and a custom stacking ensemble trained on text embeddings and features. Moreover, Hourrane and Benlahmer [37] focuses on the application of stylometry in intrinsic plagiarism detection, specifically in identifying changes in writing style. The authors propose a novel approach that combines syntactic structures, attention mechanisms, and contextualized word embeddings. They introduce a style embedding technique that utilizes syntactic trees and a pre-trained Multi-Task Deep Neural Network (MT-DNN). The model incorporates attention mechanisms and employs both a Bidirectional Long

Short-Term Memory (BiLSTM) and a Convolutional Neural Network (CNN) max pooling for sentence encoding. Summary of machine and ensemble learning based techniques is shown in Table 13.

Supervised machine learning and ensemble learning techniques have shown great promise in the field of intrinsic plagiarism detection. By using labeled data to train classifiers and leveraging the collective intelligence of multiple models through ensemble methods, these techniques have demonstrated high accuracy and effectiveness in identifying plagiarized passages within text documents. The use of various feature extraction techniques and models, such as Random Forest, SVM, and Logistic Regression, has allowed for a comprehensive analysis of writing styles and deviations, leading to improved detection performance. Moreover, the combination of different classifiers through ensemble learning has further enhanced the robustness and reliability of plagiarism detection systems.

2) UNSUPERVISED MACHINE LEARNING TECHNIQUES

Unsupervised machine learning-based techniques in intrinsic plagiarism detection employ algorithms to group similar textual segments within a document based on their stylistic or semantic features. These techniques aim to identify clusters that exhibit consistent writing style and detect any deviations that may indicate potential instances of plagiarism. By leveraging clustering algorithms such as k-means, DBSCAN, or hierarchical clustering, these techniques enable the detection of patterns and similarities within a document, assisting in the identification of plagiarized passages as shown in Table 14.

POS tagging is used to provide valuable insights into the structure and meaning of text. Authors have used POS in their plagiarism detection approaches. Such as, Zuo et al. [53] focuses on the challenging task of detecting style changes and determining the number of authors in multi-author documents. The authors present a two-module system to tackle this problem. The first module distinguishes between single-author and multi-author documents, while the second module determines the exact number of authors within the multi-author documents. The evaluation results highlight the difficulty of automated style change detection, indicating that it remains a challenging task. Also, Khan [75] focuses on the task of detecting style breaches within a single document for an unknown number of authors, as part of the PAN 2017 Author Identification challenge. The proposed model is an unsupervised approach that identifies style breaches and marks text boundaries using various stylistic features. The model utilizes a sentence window during its unsupervised analysis, which can be expanded to neighboring sentences. By combining well-known stylometric features with additional features, the model employs unsupervised classification to detect and mark passage boundaries based on style breaches.

On the other hand, few other feature extraction techniques i.e., statistical and extrinsic features are also used in clustering based approaches. Such as, Saini et al. [55] proposes an approach that utilizes stylometric features and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering to identify the writing style of authors in the document. The system provides an interactive interface where users can upload a text document and view the plagiarism detection results and analysis, including graphs. Moreover, Brooke and Hirst [56] presents an approach to intrinsic plagiarism detection that emphasizes extrinsic features over surface features. It utilizes a vector-space model and incorporates latent semantic analysis from an external corpus. The paper introduces a method to handle small and imbalanced span sizes and focuses on linguistically motivated features for poetry segmentation.

Unsupervised machine learning techniques have proven to be valuable tools for intrinsic plagiarism detection, particularly in scenarios where labeled data is scarce or unavailable. The application of clustering algorithms such as K-Means, DBSCAN, and agglomerative hierarchical clustering has allowed for the automatic segmentation of text documents based on stylistic variations, enabling the identification of potential plagiarized passages without the need for reference corpora.

3) STATISTICAL AND DISTANCE-BASED APPROACHES

Statistical and distance-based approaches are commonly used in intrinsic plagiarism detection to measure the similarity or dissimilarity between text documents. These techniques employ various statistical metrics, such as character n gram profiles, frequency distance and hashing to quantify the degree of overlap or divergence in terms of linguistic features, word frequencies, or stylistic patterns as shown in Table 15. By analyzing the statistical properties and distance measures, these approaches can effectively identify potential instances of plagiarism within a single document, even in the absence of a reference collection, providing valuable insights into the presence of copied or inconsistent text passages.

Lexical with character n-gram is a feature extraction technique commonly by authors in intrinsic plagiarism task. This approach captures the lexical patterns and similarities between documents, allowing for effective identification of plagiarized passages based on the shared character sequences. Kestemont et al. [57] presents a novel approach to intrinsic plagiarism detection. It proposes dividing a suspicious document into consecutive windows and representing them as vectors of character trigram frequencies. A distance matrix is then constructed to compare each window with others, using a modified normalized distance measure. An outlier detection algorithm based on Principal Components Analysis is applied to the distance matrix to identify plagiarized sections. Also, Kuta and Kitowski [58] aims to improve intrinsic plagiarism detection using the character n-grams profiles method. By refining the method's parameters, introducing richer feature sets, and achieving higher plagdet scores on recognized corpora, the authors make significant contributions to the field. Moreover, Stamatatos [60] presents a

TABLE 13. Summary of machine learning based techniques.

Ref	Year	Language	Feature	Techniques	Dataset	Finding
[36]	2020	English	BERT Embeddings	Random Forest	PAN 2020	F1 score of 0.86 for detecting style changes and an F1 score of 0.64 for identifying multi-author documents
[7]	2021	English	BERT Embeddings	LightGBM, Random Forest, MLP, KNN, BernoulliNB	PAN 2021	The proposed model achieves high performance with a macro-averaged F1-score of 0.7954 for single- and multi-author classification
[37]	2019	English	Constituency Trees Embedding (CTE), Dependency Trees Embedding (DTE)	MT-DNN, BiLSTM, CNN	PAN-PC-11, PAN 2018	BiLSTM achieved the best result with an accuracy of nearly 88%,
[38]	2019	English	Lexical with TF.IDF	Ensemble Learning with SVM, Random Forest, AdaBoost, MLP, and LightGBM	PAN 2018	Stacking performed well with WindowDiff of 0.5719
[39]	2014	English, Spanish, Arabic	Lexical with N gram	Naïve Bayes	PANPC-09, PANPC-11, InAra	The proposed method achieves best results in term of recall:0.69 in inAra dataset.
[67]	2015	English	Lexical with character n gram	Sequential minimal optimization	PAN-PC-11	The proposed feature set achieves an F-Score of 85.10%
[41]	2017	English	Semantic and Stylometric	SVM, Decision Tree	PAN 2009, PAN 2011	Achieved the best results in terms of F score and recall with 0.419 and 0.600 respectively.
[47]	2010	English	Stylometric features with POS and Text statistics	Neural network	PAN 2009	The proposed NN model achieved 83% on no plagiarized and 60% on plagiarized documents.
[43]	2019	English	Stylometric features with LSA	Multi-Layer Perceptron (MLP) neural networks	Corpus of English Novels (CEN)	The proposed methodology achieves the detection accuracy of 0.9715
[44]	2021	English	Stylometric features with POS and Text statistics	Logistic Regression	PAN 2021	The model achieved F1 scores of 0.634 on Task 1, 0.657 on Task 2, and 0.432 on Task 3

new method for intrinsic plagiarism detection that focuses on stylistic changes within a document. The method utilizes character n-gram profiles and a dissimilarity measure originally designed for author identification. Additionally, heuristic rules are proposed to detect plagiarism-free documents, identify plagiarized passages, and mitigate the impact of irrelevant style changes.

On the other hand, syntactic and stylometric features are also used as a feature extraction technique in different statistical and distance based approaches. For example, Vartapetiance and Gillam [62] discusses the authors' attempts at Authorship Attribution, Intrinsic Plagiarism Detection, and Sexual Predator Identification tasks. They initially explored cues of deception but found them less useful. Instead, they propose simple approaches and report their findings on detection rates using various features and techniques. Also, Tschuggnall and Specht [63] presents a novel approach for intrinsic plagiarism detection that focuses on analyzing the grammar of a suspicious document. The method involves splitting the text into sentences and calculating grammar trees. A distance matrix is created by comparing these trees using the pq-gram-distance, a variation of the tree edit distance. Suspicious sentences are identified based on significant differences in grammar and their deviation from a Gaussian normal distribution. The algorithm aims to identify potentially plagiarized sections within the document by detecting syntactical changes. Importantly, the approach relies solely on the given document and does not require a reference corpus. Moreover, Oberreuter et al. [64] proposes the intrinsic approach that focuses on outlier detection

TABLE 14. Summary of unsupervised learning based techniques.

Ref	Year	Language	Feature	Techniques	Dataset	Finding
[53]	2019	English	Token and POS distribution features.	K-Means, Agglomerative Hierarchical	PAN 2019	Accuracy of 0.6 OCI of 0.808
[75]	2017	English	Stylometric features with POS	K-Means, Agglomerative Hierarchical	PAN 2017	The proposed unsupervised model effectively detects style breaches and marks passage boundaries within a single document with WinDiff of 0.4799
[55]	2021	English	statistical stylometric features	DBSCAN	N/A	The system provides an interactive interface for users and facilitates analysis of the detected plagiarism.
[56]	2012	English	Vector-space model with extrinsic features	K-Means, Agglomerative Hierarchical	PAN 2012	Model achieved 0.969 recall in individual paragraph clustering.

TABLE 15. Summary of statistical and distance-based based techniques.

Ref	Year	Language	Feature	Techniques	Dataset	Finding
[57]	2011	English	Lexical with character n gram	PCA with distance Scores	PAN-PC-2011	The system achieves lower overall plagiarism detection score of .1679
[58]	2014	English	Lexical with character n gram	Character n-grams profiles method	PAN-PC-09, PAN-PC-11	The system achieves overall plagdet score of 33.41%
[60]	2009	English	Lexical with character n gram	Character n grams profiles method	IPAT-DC IPAT-CC	The system achieves overall score of 0.2462
[61]	2011	English	uni-grams	Frequency based algorithm	PAN 2009, PAN 2010	Proposed model achieves overall score of 0.3457.
[62]	2012	English	frequent words	Frequency Difference	PAN 2012	frequent words yielded an overall accuracy of 91.1%.
[63]	2012	English	Syntactic	pq-gram distance	PAN 2011	The proposed system achieved F score of 60%.
[64]	2011	English	Syntactic with POS tagging	Lempel-Ziv compression	PAN 2010, PAN 2011	The system achieves overall score of 0.3457
[65]	2006	English	Stylometric features	Quantification of style aspects	Manually created corpus	The proposed method achieves recall values of 85% and a precision of 75%
[12]	2018	English	statistical stylometric features	Hashing Classifier, Counting Classifier	PAN 2018	model has achieved accuracy score 0.803 on the test dataset.

to identify changes in the author's style. They include the application of outlier detection techniques to enhance intrinsic plagiarism detection based on writing style deviations. Likewise, Meyer Zu Eissen and Stein [65] focuses on identifying plagiarized passages within a single document without the need for a reference collection. The authors propose a taxonomy of plagiarism delicts and corresponding detection methods. They introduce new features for quantifying style aspects and provide a publicly available plagiarism corpus for benchmarking.

Subsequently, uni-grams and frequent words based features are also used as input by different authors. For instance, Safin and Ogaltsov [12] focuses on style change detection in the PAN'18 author identification task. The authors propose a supervised learning approach, utilizing text statistics, hashing, and high-dimensional text vectors as features. An ensemble of classifiers is employed, each independently trained on different feature groups. The preprocessing procedure varies for each classifier. Also, Oberreuter et al. [61] introduces outlier detection techniques for enhancing both intrinsic and external plagiarism detection. It utilizes self-based information algorithms for intrinsic plagiarism detection and text processing algorithms with space search reduction techniques for external plagiarism detection. The inclusion of outlier detection methodologies improves the performance of plagiarism detection.

It is evident from above discussion that statistical and distance-based approaches have played a significant role in advancing intrinsic plagiarism detection methods. By leveraging statistical properties of language, such as word frequencies and n-gram profiles, these techniques have been successful in quantifying stylistic variations within text documents and identifying potential instances of plagiarism. The use of distance measures, such as pq-gram distance and frequency difference, has allowed for the comparison of text segments, enabling the detection of deviations in writing style. Additionally, these approaches have shown promise in cross-language plagiarism detection, where they have been effective in measuring textual similarity across different languages.

4) CLASSICAL AND TRADITIONAL TECHNIQUES

In addition to the commonly used techniques in intrinsic plagiarism detection, there are several classical and traditional that have shown promise in this field as shown in Table 16. One such technique is the utilization of pre-trained encoder-decoder models, which leverage deep learning architectures to capture the underlying semantic structure and meaning of text documents. Another approach involves the use of a style change function, which identifies variations in writing style within a document and can be helpful in detecting potential instances of plagiarism. The Wilcoxon Signed Rank test, a statistical hypothesis test, can be employed to assess the significance of differences between distributions of features in original and suspicious documents. Lucene, a widely-used indexing library, can be utilized for efficient storage and retrieval of textual data, aiding in the identification and comparison of plagiarized passages. Additionally, combining multiple discriminators using voting schemes, classical discriminant analysis, and singular value decomposition techniques have shown promise in enhancing the accuracy and effectiveness of plagiarism detection algorithms by extracting meaningful features and reducing the dimensionality of the data.

The above mentioned techniques mostly used stylometric based features i.e., statistical, stop words, suffix etc. For example, Muhr et al. [69] presents a hybrid system, focusing on plagiarism detection for both translated and non-translated intrinsic plagiarized passages. Intrinsic plagiarism detection is achieved by identifying major style changes using word suffixes and a linear text segmentation algorithm. Moreover, Meyer Zu Eissen et al. [71] addresses the challenge of automatic plagiarism detection for text documents without relying on a reference collection. Authors proposes a method to identify potentially plagiarized passages within a single document by analyzing changes in writing style. They introduce new style features and presents encouraging results based on experiments conducted on a test corpus. Also, Bensalem et al. [70] presents a preliminary study on intrinsic plagiarism detection in Arabic textual documents. The authors conduct experiments to investigate the impact of language-independent stylistic features on distinguishing plagiarized and non-plagiarized Arabic text. They utilize the Stylysis tool to measure these features on a small-sized corpus. Furthermore, [48] addresses the growing issue of plagiarism by presenting an integrated approach combining Latent Semantic Indexing (LSI) and Stylometry for intrinsic plagiarism detection. LSI is used to analyze the term document matrix of the dataset, while stylometry is employed to approximate human writing style. The study includes experiments to explore the impact of the dimensionality reduction parameter in LSI and evaluates the effectiveness of the integrated approach compared to using LSI and stylometry separately. Similarly, Karaś et al. [68] presents methods for style breach detection. The proposed method involves a statistical approach based on tf-idf features is employed to characterize documents. By applying the Wilcoxon Signed Rank test to these features, the paper identifies style breaches. The submitted system for style breach detection achieved the best result in terms of F-score for WinPR, which is used to rank all participating teams.

On the other hand, lexical based and word embedding are also used by different authors as input for intrinsic plagiarism detection tasks. Such as, Safin and Kuznetsova [59] focuses on the style breach detection task and presents a method that utilizes high-dimensional vector space mapping for sentences. The approach involves using a pre-trained encoder-decoder model to generate sentence vectors that consider the context of neighboring sentences. These vectors are then used to construct an author style function and identify outliers. The method is evaluated using the PAN-2017 collection for the style breach detection task. The proposed approach utilizes neural phrase embeddings, where each sentence is mapped into a high-dimensional vector space using the skip-thoughts model. The sentence vectors capture dependencies with preceding and succeeding sentences. A similarity matrix is constructed to detect outliers among all the sentences in the document. Moreover, Kuznetsov et al. [8] focuses on intrinsic plagiarism detection and author diarization. The authors propose a method that constructs an author style function using text sentence features and detects outliers to detect plagiarism. They adapt this method for the diarization problem by segmenting author style statistics on different text parts representing different authors.

These techniques range from pre-trained encoder-decoder models to character n-gram profiles, Lempel-Ziv

TABLE 16. Summary of classical and traditional techniques.

Ref	Year	Language	Feature	Techniques	Dataset	Finding
[59]	2017	English	Word Embedding's with n-gram frequency	Pre-trained encoder decoder model	PAN 2017	The proposed method achieves a WinF measure of 0.28
[14]	2011	English	Lexical with character n gram	Style change function	PAN-PC-2011	The system achieves overall plagdet score of 0.069
[68]	2017	English, Dutch, and Greek	Stylometric features with with TF.IDF	The Wilcoxon Signed Rank test	PAN 2017	The submitted system achieved the best result in terms of F-score for WinPR.
[69]	2010	English	Stylometric features with stop words and Stem-suffix	Lucene for indexing	PAN 2010	The system achieved the third overall rank with a score of 0.6948
[70]	2012	Arabic	Language-independent stylistic features	Combining discriminators using voting schemes	Manually created corpus	Combining discriminator raise the baseline precision and recall relatively high (over 45%)
[71]	2007	English	statistical stylometric features	Classical discriminant analysis	Manually created corpus	Average Frequency Class based Wilks Lambda achieves 0.723 score.
[48]	2013	English	Stylometric features with LSA	Singular Value Decomposition	MED collection	Function Honore (R) has achieved better in terms of precision and recall with 40% DR

compression, and style change functions. Each of these methods brings unique insights and contributions to the field, offering alternative ways to capture and analyze stylistic variations within text documents. The use of pre-trained models and deep learning techniques has shown promise in learning complex patterns of writing style, while character n-grams have proven effective in capturing fine-grained linguistic details. Additionally, the incorporation of compression and style change functions has allowed for novel ways of detecting plagiarism and identifying segments with distinct writing styles.

V. EVALUATION METRICS

The evaluation of intrinsic plagiarism detection techniques relies on a set of diverse metrics that provide insights into the performance of these methods [85]. These metrics encompass classification-based measures, clustering evaluation criteria, as well as metrics that are specifically designed for plagiarism detection tasks [58] as shown in figure 6. Among the commonly used classification metrics are Accuracy, Precision, Recall, and F Score. Accuracy quantifies the overall correctness of predictions, while Precision measures the proportion of true positives among predicted positives. Recall, on the other hand, captures the proportion of true positives among actual positives, and F Score balances Precision and Recall for imbalanced datasets.

In the context of clustering methods, Granularity is a metric that assesses the level of detail in the clustering results. It indicates how fine-grained the clusters are. Additionally, PlagDet, short for Plagiarism Detection, is an evaluation criterion that specifically evaluates the performance of clustering methods in the context of plagiarism detection.

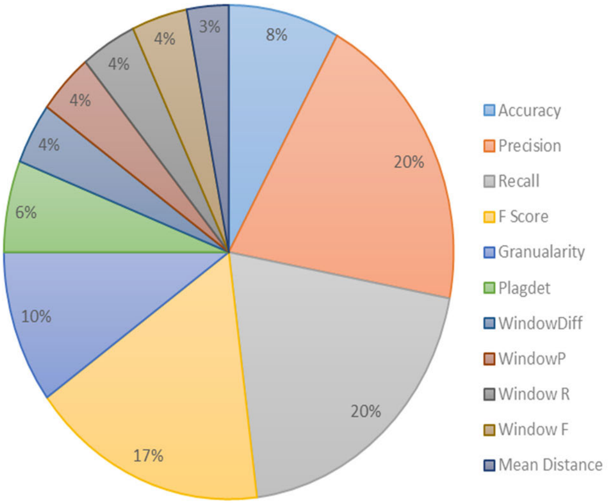

FIGURE 6. Distribution of evaluation metrics in papers.

Furthermore, there are metrics like WindowDiff, WindowP, WindowR, and WindowF that are tailored for comparing clustering partitions. These metrics help measure the quality of clusters by evaluating the probability of data points belonging to the same cluster (WindowP) or different clusters (WindowR), along with a harmonic mean (WindowF) of these probabilities. WindowDiff measures the difference between two clustering partitions. Lastly, Mean Distance is a statistical metric that gauges the average distance between data points within clusters. This metric helps assess the compactness of clusters.

VI. APPLICATIONS

Intrinsic plagiarism detection techniques have found applications in various domains beyond traditional plagiarism detection scenarios as shown in Figure 7. These techniques offer valuable insights and contribute to diverse areas where maintaining originality and authenticity of textual content is crucial. The following sections highlight some of these applications.

A. EDUCATIONAL SETTINGS AND AI CONTENT GENERATION

Educational settings have long relied on intrinsic plagiarism detection to uphold the originality of student assignments and essays, ensuring academic integrity within educational institutions [58]. In today's evolving landscape, where AI content generation tools like ChatGPT are becoming increasingly sophisticated, the role of intrinsic plagiarism detection is more critical than ever [3]. These tools can generate highly convincing academic content, posing a potential challenge to the verification of student work. Similarly, plagiarism detection in research papers, especially as AI-generated content becomes prevalent, safeguards the credibility of scientific literature, maintaining the highest academic standards [86]. In the context of higher education, theses and dissertations are subject to plagiarism detection not only to authenticate the originality of advanced academic work but also to ensure that AI-generated content doesn't compromise the rigor of scholarly contributions [87]. Thus, intrinsic plagiarism detection remains an essential safeguard against the inadvertent or intentional use of AI-generated content in academic and research contexts.

B. LITERARY AND CREATIVE WORKS

Intrinsic plagiarism detection extends its influence to literary and creative works, guaranteeing the authenticity of novels and stories by protecting authors' creative expressions [88]. The detection of copied poetry and lyrics further preserves the artistic contributions of poets and songwriters [88].

C. LEGAL AND INTELLECTUAL PROPERTY

In legal and intellectual property domains, intrinsic plagiarism detection assumes a crucial role in preventing the submission of duplicate patent claims, ensuring the uniqueness of intellectual property [89]. Furthermore, identifying copyright infringement in books, articles, and digital media safeguards the rights of content creators [90].

D. JOURNALISM AND NEWS MEDIA

Journalism and news media benefit from intrinsic plagiarism detection by maintaining journalistic integrity, ensuring originality in news reporting [91]. This technology also aids in upholding the credibility of journalists and columnists by detecting copied content in opinion pieces and columns [92].

E. CONTENT CREATION AND BLOGGING

In the digital age, content creation and blogging rely on intrinsic plagiarism detection to preserve originality in online articles, blogs, and web-based content [93].

F. SOCIAL MEDIA AND ONLINE PLATFORMS

Similarly, social media and online platforms leverage this technology to prevent the proliferation of copied content across various digital platforms [94].

G. TRANSLATION AND LANGUAGE SERVICES

Translation and language services employ intrinsic plagiarism detection to validate the originality of translated texts, preserving the integrity of the translation process [95].

H. SOFTWARE AND CODE DEVELOPMENT

Furthermore, in software and code development, the technology plays a pivotal role in detecting plagiarism in programming code and algorithms, ensuring the integrity of software development [96]. Software documentation's authenticity is also maintained through intrinsic plagiarism detection, thereby contributing to the credibility of technical guides [96].

VII. EVOLUTION, CHALLENGES AND WAY FORWARD

The progression of intrinsic plagiarism detection techniques has unfolded a dynamic landscape, punctuated by innovations and challenges. This section delves into the evolution of these techniques, tracking their journey over time and highlighting pivotal developments that have shaped the field. Moreover, it sheds light on the persistent challenges that have accompanied this evolution, ranging from the complexities of dealing with low-resource languages to the subtleties of identifying sophisticated forms of plagiarism. As the field advances, a forward-looking perspective is also presented, exploring potential pathways to address these challenges and further enhance the effectiveness and robustness of intrinsic plagiarism detection methods.

A. EVOLUTION

Intrinsic plagiarism detection methods have evolved significantly over the years, employing a range of techniques to tackle the challenges in identifying stylistic changes and potential plagiarism within a document as shown in Figure 8. The early years of research focused on quantifying style aspects [65] and utilizing classical discriminant analysis [71] to distinguish between original and plagiarized content. In 2009, the character n-grams profiles method [60] emerged, utilizing character-based n-gram features for detection. Subsequently, the adoption of neural networks [42] in 2010 brought new capabilities for modeling and detecting intricate stylistic variations.

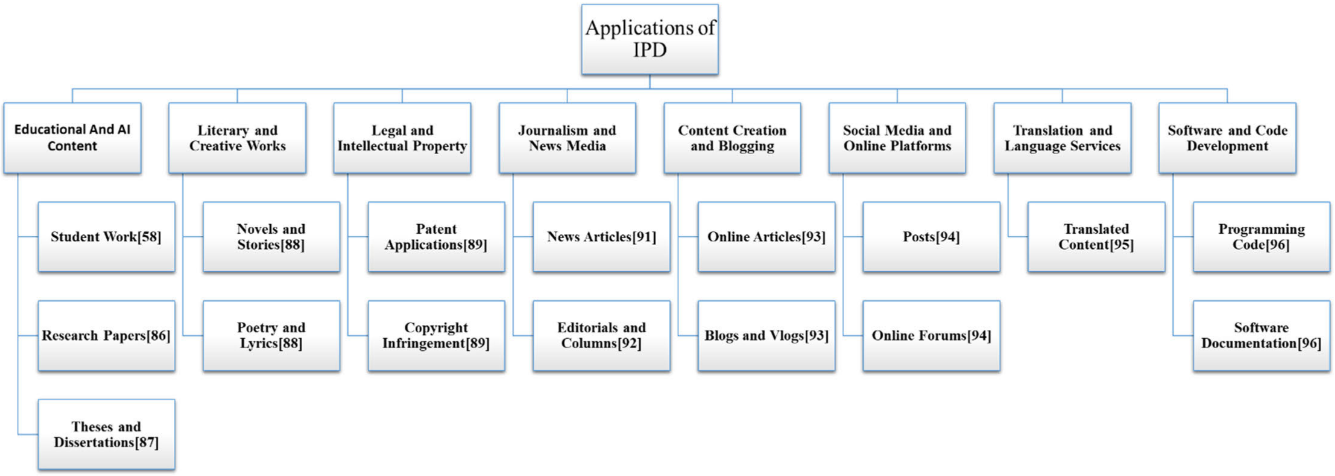
Applications of IPD
Educational And AI Content
Student Work[58]
Research Papers[86]
Theses and Dissertations[87]
Literary and Creative Works
Novels and Stories[88]
Poetry and Lyrics[88]
Legal and Intellectual Property
Patent Applications[89]
Copyright Infringement[89]
Journalism and News Media
News Articles[91]
Editorials and Columns[92]
Content Creation and Blogging
Online Articles[93]
Blogs and Vlogs[93]
Social Media and Online Platforms
Posts[94]
Online Forums[94]
Translation and Language Services
Translated Content[95]
Software and Code Development
Programming Code[96]
Software Documentation[96]

FIGURE 7. Applications of intrinsic plagiarism detection techniques.

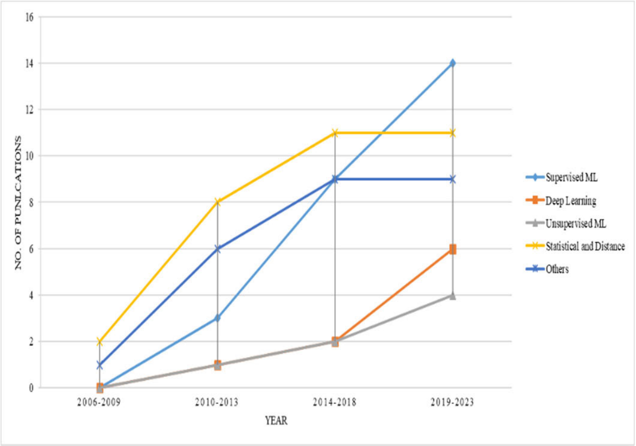

FIGURE 8. Evolution of IPD techniques over the years.

Lucene for indexing [69] was introduced in 2010, offering efficient information retrieval techniques for intrinsic plagiarism detection. In the same year, PCA with distance scores [57] utilized principal component analysis to capture stylistic differences based on distance metrics. Frequency-based algorithms [61] gained attention in 2011, focusing on the analysis of word and phrase frequencies to identify potential plagiarism.

Advancements continued with the adoption of Lempel-Ziv compression [64] in 2011, leveraging compression algorithms to detect stylistic changes and irregularities. Style change functions [14] were proposed in the same year to identify shifts in writing style within a document. Clustering techniques such as K-Means and Agglomerative Hierarchical clustering [56] were employed in 2012 to group similar text segments based on stylistic similarities.

Frequency difference [62] and pq-gram distance [63] approaches in 2012 emphasized the comparison of frequency distributions and n-gram patterns for plagiarism detection. The concept of combining discriminators using voting schemes [70] was introduced, harnessing the collective decision-making of multiple models in 2012.

Singular Value Decomposition (SVD) [72] emerged in 2013, enabling the decomposition of stylistic features into latent components for analysis. Naïve Bayes [39] in 2014 brought probabilistic modeling for plagiarism detection based on stylometric features. Character n-grams profiles method [58] further advanced in 2014, employing character-based n-gram profiles for detecting plagiarism.

In subsequent years, techniques such as Sequential Minimal Optimization [40], Support Vector Machines (SVM) and Decision Trees [41], Pre-trained encoder-decoder models [67], and the Wilcoxon Signed Rank test [68] made notable contributions to the field. Hashing classifiers and counting classifiers [38] were introduced in 2018 for efficient feature representation. MT-DNN with BiLSTM and CNN [37] emerged in 2019, harnessing multi-task deep neural networks for style change detection. Ensemble learning with SVM, Random Forest, AdaBoost, MLP, and LightGBM [38] showcased the benefits of combining multiple models for improved performance. Recent advancements include the utilization of Logistic Regression [44], DBSCAN [55] and the application of Random Forest [36] LightGBM, Random Forest, MLP, KNN, and BernoulliNB [7] in 2021.

B. RESEARCH GAPS AND CHALLENGES

Intrinsic plagiarism detection poses several research gaps and challenges that researchers and practitioners must overcome to effectively identify and address instances of plagiarism within a given document.

- **Absence of Reference Collection:** One significant challenge is the absence of a reference collection or known sources for comparison. Unlike extrinsic plagiarism detection, where documents are compared against external sources, intrinsic detection relies solely on the analysis of a single document. This lack of external references makes it difficult to distinguish genuine stylistic changes from instances of plagiarism, especially when the plagiarized passages come from non-digital or unavailable sources [65].

- **Variability of Writing Styles:** Another challenge is the variability and complexity of writing styles. Authors may employ different writing styles intentionally or unintentionally, leading to variations within their own work. These variations can be caused by factors such as changes in topic, writing intent, or audience. Distinguishing genuine style changes from potential plagiarism requires robust techniques that can capture and quantify subtle stylistic nuances and differentiate them from intentional or natural variations [48].
- **Optimal Feature Selection:** The identification of optimal features and their representation is another challenge in intrinsic plagiarism detection. Different features, such as character n-grams, word frequencies, syntactic structures, or semantic patterns, have been explored to capture and quantify writing style. However, selecting the most informative and discriminative features for detection is a complex task. Additionally, representing these features effectively to capture the unique characteristics of a document and its potential style changes requires careful consideration.
- **Small-Scale or Partial Plagiarism:** The detection of small-scale or partial plagiarism presents another challenge. Intrinsic plagiarism can involve plagiarized sections that are fragmented or dispersed throughout a document. Identifying these instances of partial plagiarism, where only specific sentences or phrases have been copied, requires advanced algorithms that can pinpoint subtle similarities and differences within the document's textual content [97].
- **Evolving Techniques and Technologies:** Furthermore, the evolving nature of plagiarism techniques and the continuous advancements in language generation and manipulation technologies pose ongoing challenges for intrinsic plagiarism detection. Plagiarists may employ sophisticated methods to obfuscate or alter their writing styles, making it increasingly difficult to detect instances of plagiarism solely based on intrinsic characteristics [1].
- **Maintaining Effectiveness Amidst AI-Generated Content**: As AI content generation tools like ChatGPT continue to advance, one significant challenge in intrinsic plagiarism detection is maintaining its effectiveness in identifying instances of plagiarism in AI-generated content. These AI systems can produce human-like text, making it difficult to distinguish between genuinely authored content and AI-generated material [3].

1) LIMITATIONS IN LOW RESOURCE LANGUAGES

Intrinsic plagiarism detection in low-resource languages poses unique challenges compared to high-resource languages.

- **Scarcity of Linguistic Resources:** One major challenge is the scarcity of resources such as labeled datasets, linguistic tools, and language models specific to low-resource languages. High-resource languages benefit from extensive research and development, resulting in abundant linguistic resources, well-established language models, and sophisticated tools for text analysis. In contrast, low-resource languages often lack comprehensive linguistic resources and pre-trained models, making it difficult to apply state-of-the-art techniques directly [98]
- **Limited Availability of Reference Materials:** Another challenge is the limited availability of reference materials and corpora for low-resource languages. In high-resource languages, researchers can draw upon vast collections of texts, books, and online sources for reference and comparison purposes. However, in low-resource languages, such resources may be scarce, incomplete, or not readily accessible in digital form. This limitation hinders the ability to accurately identify and differentiate between genuine style variations and instances of plagiarism in low-resource languages [98].
- **Lack of Linguistic Diversity:** The lack of linguistic diversity and variability within low-resource languages presents another challenge. High-resource languages encompass a wide range of dialects, registers, and writing styles, which contribute to the complexity and diversity of the language. Intrinsic plagiarism detection algorithms can leverage these variations to identify stylistic changes. In contrast, low-resource languages often exhibit limited linguistic variation, with fewer dialects or registers. This reduced variability makes it more challenging to distinguish genuine style changes from potential plagiarism in low-resource languages [99].
- **Scarcity of Language-Specific Features:** Moreover, the scarcity of language-specific stylometric features further complicates intrinsic plagiarism detection in low-resource languages. Stylometric features, such as n-grams, word frequencies, or syntactic patterns, play a crucial role in capturing and quantifying writing style. However, these features may not be as effective in low-resource languages due to limited linguistic resources and variations. Developing language-specific stylometric features tailored to the unique characteristics of low-resource languages is a critical challenge [100].

C. WAY FORWARD

In tackling the challenges of intrinsic plagiarism detection, several promising directions can guide future research and development.

- **Transfer Learning for Low-Resource Languages:** One avenue is the exploration of transfer learning techniques, where knowledge gained from high-resource languages can be adapted and applied to low-resource languages [101]. Leveraging pre-trained models and linguistic resources from high-resource languages could mitigate the scarcity of resources and enhance detection accuracy in low-resource language contexts.

- **Collaboration and Resource Digitization:** Additionally, the collaboration between researchers, linguistic experts, and language preservation initiatives is pivotal. Efforts to digitize and create language-specific resources can substantially support intrinsic plagiarism detection in low-resource languages. Developing open-access corpora, linguistic tools, and language models for these languages can significantly alleviate the challenge of limited resources [102].
- **Unsupervised and Self-Supervised Learning Advancements:** Furthermore, the advancement of unsupervised and self-supervised learning methods is promising for addressing the lack of labeled data. These techniques can reduce dependency on extensive annotated datasets, enabling intrinsic plagiarism detection to operate effectively even with limited training samples [72].
- **Cross-Lingual Transfer Techniques for Linguistic Variability:** To address the challenge of linguistic variability in low-resource languages, researchers could investigate cross-lingual transfer techniques. These techniques aim to capture stylistic variations by leveraging similarities and differences across languages, potentially compensating for the reduced diversity within a single language [69].
- **Developing Language-Agnostic Stylometric Features:** Developing language-agnostic stylometric features could help mitigate the scarcity of language-specific features for low-resource languages. By focusing on universal stylistic elements that transcend language boundaries, researchers can create effective feature representations for intrinsic plagiarism detection across diverse language contexts [55].

VIII. CONCLUSION

This systematic literature review (SLR) has provided a comprehensive overview of the field of Intrinsic Plagiarism Detection (IPD). We explored various aspects of IPD, including common datasets, preprocessing and feature extraction techniques, detection methods, evaluation metrics, and applications. Through this review, it became evident that IPD is a critical area of research with wide-ranging applications in academia, literature, legal domains, journalism, content creation, and more.

This SLR also identified the evolution of IPD techniques, from traditional methods to advanced approaches that leverage machine learning and natural language processing. Moreover, significant challenges faced by researchers and practitioners in IPD, including the absence of reference collections, variability in writing styles, optimal feature selection, and the detection of small-scale plagiarism are also discussed. One notable aspect of this SLR is the identification of the unique set of challenges faced when dealing with low-resource languages in the context of IPD. These languages often lack comprehensive linguistic resources, well-established language models, and large-scale datasets. As a result, the development and application of IPD techniques in low-resource languages are hindered.

Moving forward, IPD holds immense potential for further advancements, especially with the continuous development of AI and NLP technologies. Researchers should continue to address the challenges posed by IPD and work towards enhancing the accuracy and efficiency of plagiarism detection methods, with a particular focus on strategies for low-resource languages. Additionally, the utilization of IPD in various domains, such as journalism, academia, and creative writing, should be explored further.

REFERENCES

[1] T. Foltýnek, N. Meuschke, and B. Gipp, "Academic plagiarism detection: A systematic literature review," *ACM Comput. Surv.*, vol. 52, no. 6, pp. 1–42, Nov. 2020, doi: 10.1145/3345317.

[2] V. A. Oloo, C. Otieno, and L. A. Wanzare, "A literature survey on writing style change detection based on machine learning: State-of-the-art–review," *Int. J. Comput. Trends Technol.*, vol. 70, no. 5, pp. 15–32, May 2022, doi: 10.14445/22312803/ijctt-v70i5p103.

[3] Y. K. Dwivedi et al., "Opinion paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *Int. J. Inf. Manage.*, vol. 71, Aug. 2023, Art. no. 102642, doi: 10.1016/j.ijinfomgt.2023.102642.

[4] M. Khalil and E. Er, "Will ChatGPT get you caught? Rethinking of plagiarism detection," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2023, pp. 1–13, doi: 10.1007/978-3-031-34411-4_32.

[5] Y. Kim, "The pilot study in qualitative inquiry: Identifying issues and learning lessons for culturally competent research," *Qualitative Social Work*, vol. 10, no. 2, pp. 190–206, Jun. 2011, doi: 10.1177/1473325010362001.

[6] S. Alzahrani and H. Aljuaid, "Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1110–1123, Apr. 2022, doi: 10.1016/j.jksuci.2020.04.009.

[7] E. Strøm, "Multi-label style change detection by solving a binary classification problem," in *Proc. CEUR Workshop*, vol. 2936, 2021, pp. 2146–2157.

[8] M. P. Kuznetsov, A. Motrenko, R. Kuznetsova, and V. V Strijov, "Methods for intrinsic plagiarism detection and author Diarization.," in *Proc. CLEF Work. Notes*, 2016, pp. 912–919.

[9] A. Misargopoulos, F. Nikolopoulos-Gkamatsis, K. Nestorakis, A. Tzoumas, G. Giannakopoulos, C.-A. Gizelis, and M. Kefalogiannis, "Building a knowledge-intensive, intent-lean, question answering chatbot in the telecom industry—Challenges and solutions," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, vol. 652. Springer, 2022.

[10] L. Suanmali, N. Salim, and M. S. Binwahlan, "Fuzzy logic based method for improving text summarization," vol. 2, no. 1, 2009, *arXiv:0906.4690*.

[11] L. Suanmali, M. S. Binwahlan, and N. Salim, "Sentence features fusion for text summarization using fuzzy logic," in *Proc. 9th Int. Conf. Hybrid Intell. Syst. HIS*, vol. 1, 2009, pp. 142–146, doi: 10.1109/HIS.2009.36.

[12] K. Safin and A. Ogaltsov, "Detecting a change of style using text statistics: Notebook for PAN at CLEF 2018," in *Proc. CEUR Workshop*, vol. 2125, 2018.

[13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.

[14] S. Rao, P. Gupta, K. Singhal, and P. Majumder, "External & intrinsic plagiarism detection: VSM & discourse markers based approach notebook for PAN at CLEF 2011," in *Proc. CEUR Workshop*, vol. 1177, 2011, pp. 2–6.

[15] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on Twitter aspect-based sentiment analysis," *Appl. Intell.*, vol. 48, no. 5, pp. 1218–1232, 2018, doi: 10.1007/s10489-017-1098-6.

[16] H. Zhang and J. Wang, "Semantic WordRank: Generating finer single-document summarizations," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11314, 2018, pp. 398–409, doi: 10.1007/978-3-030-03493-1_42.

[17] A. Mansouri, L. S. Affendey, and A. Mamat, "Named entity recognition approaches," *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 2, pp. 339–344, 2008.

[18] P. Clough, "Old and new challenges in automatic plagiarism detection," *Natl. Plagiarism Advis. Serv.*, vol. 41, pp. 391–407, Feb. 2003.

[19] N. Mukhtar and M. A. Khan, "Effective lexicon-based approach for Urdu sentiment analysis," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2521–2548, Apr. 2020, doi: 10.1007/s10462-019-09740-5.

[20] M. Jiffriya, M. A. Jahan, and R. G. Ragel, "Plagiarism detection tools and techniques: A comprehensive survey," *J. Sci.*, vol. 2, no. 2, pp. 47–64, 2021.

[21] M. Ishaq, A. Abid, M. S. Farooq, M. F. Manzoor, U. Farooq, K. Abid, and M. A. Helou, "Advances in database systems education: Methods, tools, curricula, and way forward," in *Education and Information Technologies* vol. 28, no. 3. Springer, 2023.

[22] U. Farooq, M. S. M. Rahim, N. Sabir, A. Hussain, and A. Abid, "Advances in machine translation for sign language: Approaches, limitations, and challenges," *Neural Comput. Appl.*, vol. 33, no. 21, pp. 14357–14399, Nov. 2021.

[23] A. Abid, N. Hussain, K. Abid, F. Ahmad, M. S. Farooq, U. Farooq, S. A. Khan, Y. D. Khan, M. A. Naeem, and N. Sabir, "A survey on search results diversification techniques," *Neural Comput. Appl.*, vol. 27, no. 5, pp. 1207–1229, Jul. 2016.

[24] M. Ramzan, A. Abid, H. U. Khan, S. M. Awan, A. Ismail, M. Ahmed, M. Ilyas, and A. Mahmood, "A review on state-of-the-art violence detection techniques," *IEEE Access*, vol. 7, pp. 107560–107575, 2019.

[25] R. Tehseen, M. S. Farooq, and A. Abid, "Earthquake prediction using expert systems: A systematic mapping study," *Sustainability*, vol. 12, no. 6, p. 2420, Mar. 2020.

[26] M. Aria and C. Cuccurullo, "Bibliometrix : An R-tool for comprehensive science mapping analysis," *J. Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017.

[27] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "SciMAT: A new science mapping analysis software tool," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 8, pp. 1609–1630, Aug. 2012.

[28] T. A. E. Eisa, N. Salim, and S. Alzahrani, "Existing plagiarism detection techniques: A systematic mapping of the scholarly literature," *Online Inf. Rev.*, vol. 39, no. 3, pp. 383–400, Jun. 2015, doi: 10.1108/oir-12-2014-0315.

[29] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, and D. R. I. M. Setiadi, "Review of automatic text summarization techniques & methods," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1029–1046, Apr. 2022, doi: 10.1016/j.jksuci.2020.05.006.

[30] M. Sabeeh and F. Khaled, "Plagiarism detection methods and tools: An overview," *Iraqi J. Sci.*, vol. 62, no. 8, pp. 2771–2783, Aug. 2021, doi: 10.24996/ijs.2021.62.8.30.

[31] E. Stamatatos, "Authorship verification: A review of recent advances," *Res. Comput. Sci.*, vol. 123, no. 1, pp. 9–25, Dec. 2016, doi: 10.13053/rcs-123-1-1.

[32] M. Khonji, Y. Iraqi, and L. Mekouar, "Authorship identification of electronic texts," *IEEE Access*, vol. 9, pp. 101124–101146, 2021, doi: 10.1109/ACCESS.2021.3098192.

[33] F. Elberzhager, J. Münch, and V. T. N. Nha, "A systematic mapping study on the combination of static and dynamic quality assurance techniques," *Inf. Softw. Technol.*, vol. 54, no. 1, pp. 1–15, Jan. 2012.

[34] A. P. Widyassari, A. Affandy, E. Noersasongko, A. Z. Fanani, A. Syukur, and R. S. Basuki, "Literature review of automatic text summarization: Research trend, dataset and method," in *Proc. Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Jul. 2019, pp. 491–496, doi: 10.1109/ICOIACT46704.2019.8938454.

[35] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[36] A. Iyer and S. Vosoughi, "Style change detection using BERT notebook for PAN at CLEF 2020," in *Proc. CEUR Workshop*, vol. 2696, Sep. 2020, pp. 22–25.

[37] O. Hourrane and E. Habib, "Rich style embedding for intrinsic plagiarism detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 11, pp. 646–651, 2019, doi: 10.14569/ijacsa.2019.0101185.

[38] 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.

[39] I. Bensalem, P. Rosso, and S. Chikhi, "Intrinsic plagiarism detection using N-gram classes," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1459–1464, doi: 10.3115/v1/d14-1153.

[40] G. Oberreuter and J. D. Velásquez, "Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style," *Expert Syst. Appl.*, vol. 40, no. 9, pp. 3756–3763, Jul. 2013, doi: 10.1016/j.eswa.2012.12.082.

[41] A. Polydouri, G. Siolas, and A. Stafylopatis, "Intrinsic plagiarism detection with feature-rich imbalanced dataset learning," in *18th Int. Conf.Eng. Appl. Neural Netw.*, vol. 2, Athens, Greece, 2017, pp. 87–98, doi: 10.1007/978-3-319-65172-9.

[42] J. Ferrero, F. Agnes, L. Besacier, and D. Schwab, "UsingWord embedding for cross-language plagiarism detection," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, vol. 2, 2017, pp. 415–421, doi: 10.18653/v1/e17-2066.

[43] M. AlSallal, R. Iqbal, V. Palade, S. Amin, and V. Chang, "An integrated approach for intrinsic plagiarism detection," *Future Gener. Comput. Syst.*, vol. 96, pp. 700–712, Jul. 2019, doi: 10.1016/j.future.2017.11.023.

[44] R. Singh, J. Weerasinghe, and R. Greenstadt, "Writing style change detection on multi-author documents," in *Proc. CEUR Workshop*, vol. 2936, 2021, pp. 2137–2145.

[45] X. Hua, S. Li, P. Li, and Q. Zhu, "Research on intrinsic plagiarism detection resolution: A supervised learning approach," in *Proc. Workshop Chin. Lexical Semantics*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 7717, 2013, pp. 58–63, doi: 10.1007/978-3-642-36337-5_7.

[46] B. Stein, N. Lipka, and P. Prettenhofer, "Intrinsic plagiarism analysis," *Lang. Resour. Eval.*, vol. 45, no. 1, pp. 63–82, Mar. 2011, doi: 10.1007/s10579-010-9115-y.

[47] D. Curran, "An evolutionary neural network approach to intrinsic plagiarism detection," in *Proc. 20th Irish Conf. Artif. Intell. Cogn. Sci. (AICS)*, Dublin, Ireland, 2010, pp. 33–40.

[48] M. Alsallal, R. Iqbal, S. Amin, and A. James, "Intrinsic plagiarism detection using latent semantic indexing and stylometry," in *Proc. 6th Int. Conf. Develop. eSyst. Eng.*, Dec. 2013, pp. 145–150, doi: 10.1109/DeSE.2013.34.

[49] M. AlSallal, R. Iqbal, V. Palade, S. Amin, and V. Chang, "An integrated approach for intrinsic plagiarism detection," *Future Gener. Comput. Syst.*, vol. 96, pp. 700–712, Jul. 2019.

[50] J. Tian and M. Lan, "ECNU at SemEval-2016 task 1: Leveraging word embedding from macro and micro views to boost performance for semantic textual similarity," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 621–627, doi: 10.18653/v1/s16-1094.

[51] D. Ataman, J. G. C. De Souza, M. Turchi, and M. Negri, "FBK HLT-MT at SemEval-2016 task 1: Cross-lingual semantic similarity measurement using quality estimation features and compositional bilingual word embeddings," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval-)*, 2016, pp. 570–576, doi: 10.18653/v1/s16-1086.

[52] M. R. Ghaeini, "Intrinsic author identification using modified weighted KNN: Notebook for PAN at CLEF 2013," in *Proc. CEUR Workshop*, vol. 1179, 2013.

[53] C. Zuo, Y. Zhao, and R. Banerjee, "Style change detection with feed-forward neural networks notebook for PAN at CLEF 2019," in *Proc. CEUR Workshop*, vol. 2380, Sep. 2019, pp. 9–12.

[54] H. He, J. Wieting, K. Gimpel, J. Rao, and J. Lin, "UMD-TTIC-UW at SemEval-2016 task 1: Attention-based multi-perspective convolutional neural networks for textual similarity measurement," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval)*, 2016, pp. 1103–1108, doi: 10.18653/v1/s16-1170.

[55] A. Saini, M. R. Sri, and M. Thakur, "Intrinsic plagiarism detection system using stylometric features and DBSCAN," in *Proc. Int. Conf. Comput., Commun., Intell. Syst. (ICCCIS)*, Feb. 2021, pp. 13–18, doi: 10.1109/icccis51004.2021.9397187.

[56] J. Brooke and G. Hirst, ‘‘Paragraph clustering for intrinsic plagiarism detection using a stylistic vector space model with extrinsic Features,’’ in *Proc. CLEF Online Work. Notes/Labs/Workshop*, 2012, pp. 1–9. [Online]. Available: http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-BrookeEt2012.pdf

[57] M. Kestemont, K. Luyckx, and W. Daelemans, ‘‘Intrinsic plagiarism detection using character trigram distance scores,’’ in *Proc. PAN*, vol. 63, 2011, pp. 1–9.

[58] M. Kuta and J. Kitowski, ‘‘Optimisation of character n-gram profiles method for intrinsic plagiarism detection,’’ in *Proc. Int. Conf. Artif. Intell. Soft Comput.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 8468, 2014, pp. 500–511, doi: 10.1007/978-3-319-07176-3_44.

[59] K. Safin and R. Kuznetsova, ‘‘Style breach detection with neural sentence embeddings: Notebook for PAN at CLEF 2017,’’ in *Proc. CEUR Workshop*, vol. 1866, 2017.

[60] E. Stamatatos, ‘‘Intrinsic plagiarism detection using character n-gram profiles,’’ *Threshold*, vol. 2, no. 1, p. 500, 2009.

[61] G. Oberreuter, G. L’Huillier, S. A. Ríos, and J. D. Velásquez, ‘‘Outlier-based approaches for intrinsic and external plagiarism detection,’’ in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 6882, 2011, pp. 11–20, doi: 10.1007/978-3-642-23863-5_2.

[62] A. Vartapetiance and L. Gillam, ‘‘Quite simple approaches for authorship attribution, intrinsic plagiarism detection and sexual predator identification—Notebook for PAN at CLEF 2012,’’ in *Proc. Work. Notes Pap. CLEF Eval. Labs*, 2012, pp. 1–12. [Online]. Available: http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-VartapetianceEt2012.pdf

[63] M. Tschuggnall and G. Specht, ‘‘Plag-inn: Intrinsic plagiarism detection using grammar trees,’’ in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 7337, 2012, pp. 284–289, doi: 10.1007/978-3-642-31178-9_35.

[64] G. Oberreuter, G. L’Huillier, S. A. Ríos, and J. D. Velásquez, ‘‘Approaches for intrinsic and external plagiarism detection,’’ *Proc. PAN*, vol. 4, no. 5, p. 63, 2011.

[65] S. M. Zu Eissen and B. Stein, ‘‘Intrinsic plagiarism detection,’’ in *Proc. Eur. Conf. Inf. Retr.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 3936, 2006, pp. 565–569, doi: 10.1007/11735106_66.

[66] A. Magooda, A. Y. Mahgoub, M. Rashwan, M. B. Fayek, and H. Raafat, ‘‘RDI system for extrinsic plagiarism detection (RDI_RED),’’ in *Proc. Work. Forum Inf. Retr. Eval.*, vol. 1587, May 2016, pp. 126–128. [Online]. Available: http://ceur-ws.org/Vol-1587/T5-3.pdf

[67] R. Rahman, ‘‘Information theoretical and statistical features for intrinsic plagiarism detection,’’ in *Proc. 16th Annu. Meeting Special Interest Group Discourse Dialogue*, Sep. 2015, pp. 144–148, doi: 10.18653/v1/w15-4619.

[68] D. Karaś, M. Śpiewak, and P. Sobecki, ‘‘OPI-JSA at CLEF 2017: Author clustering and style breach detection: Notebook for PAN at CLEF 2017,’’ in *Proc. CEUR Workshop*, vol. 1866, 2017.

[69] M. Muhr, R. Kern, M. Zechner, and M. Granitzer, ‘‘External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system,’’ in *Proc. Notebook Papers CLEF LABs Workshops*, 2010, p. 22.

[70] I. Bensalem, P. Rosso, and S. Chikhi, ‘‘Intrinsic plagiarism detection in Arabic text: Preliminary experiments,’’ in *Proc. 2nd Spanish Conf. Inf. Retr. (CERI)*, 2012.

[71] S. M. Z. Eissen, B. Stein, and M. Kulig, ‘‘Plagiarism detection without reference collections,’’ in *Proc. 30 th Annu. Conf. Gesellschaft für Klassifikation eV Adv. Data Anal.* (Studies in Classification, Data Analysis, and Knowledge Organization), Freie Universitüt Berlin, 2007, pp. 359–366, doi: 10.1007/978-3-540-70981-7_40.

[72] A. Polydouri, E. Vathi, G. Siolas, and A. Stafylopatis, ‘‘An efficient classification approach in imbalanced datasets for intrinsic plagiarism detection,’’ *Evolving Syst.*, vol. 11, no. 3, pp. 503–515, Sep. 2020, doi: 10.1007/s12530-018-9232-1.

[73] G. Specht and M. Tschuggnall, ‘‘using grammar-profiles to intrinsically expose plagiarism in text documents,’’ in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.*, Aug. 2015, pp. 297–302.

[74] M. Tschuggnall and G. Specht, ‘‘Detecting plagiarism in text documents through grammar-analysis of authors,’’ *Datenbanksysteme für Business, Technol. und Web*, vol. 14, pp. 241–260, 2013.

[75] J. A. Khan, ‘‘Style breach detection: An unsupervised detection model: Notebook for PAN at CLEF 2017,’’ in *Proc. CEUR Workshop*, vol. 1866, 2017.

[76] Å. Rinnan, L. Nørgaard, F. van den Berg, J. Thygesen, R. Bro, and S. B. Engelsen, ‘‘Data Pre-processing,’’ in *Infrared Spectroscopy for Food Quality Analysis and Control*. San Diego, CA, USA: Academic, pp. 29–50, Oct. 2022, doi: 10.1016/B978-0-12-374136-3.00002-X.

[77] J. Su, S. Yu, and D. Luo, ‘‘Enhancing aspect-based sentiment analysis with capsule network,’’ *IEEE Access*, vol. 8, pp. 100551–100561, 2020, doi: 10.1109/ACCESS.2020.2997675.

[78] M. Abdolahi and M. Zahedh, ‘‘Sentence matrix normalization using most likely n-grams vector,’’ in *Proc. IEEE 4th Int. Conf. Knowl.-Based Eng. Innov. (KBEI)*, Dec. 2017, pp. 0040–0045, doi: 10.1109/KBEI.2017.8325018.

[79] J. R. Méndez, E. L. Iglesias, F. Fdez-Riverola, F. Díaz, and J. M. Corchado, ‘‘Tokenising, stemming and stopword removal on anti-spam filtering domain,’’ in *Proc. Conf. Spanish Assoc. Artif. Intell.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 4177, 2006, pp. 449–458, doi: 10.1007/11881216_47.

[80] M. A. Tayal, M. M. Raghuwanshi, and L. G. Malik, ‘‘ATSSC: Development of an approach based on soft computing for text summarization,’’ *Comput. Speech Lang.*, vol. 41, pp. 214–235, Jan. 2017, doi: 10.1016/j.csl.2016.07.002.

[81] J. Wang, J. Liu, and C. Wang, ‘‘Keyword extraction based on pagerank,’’ in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 4426, 2007, pp. 857–864, doi: 10.1007/978-3-540-71701-0_95.

[82] M. M. Haider, Md. A. Hossin, H. R. Mahi, and H. Arif, ‘‘Automatic text summarization using gensim Word2 Vec and K-means clustering algorithm,’’ in *Proc. IEEE Region Symp. (TENSYMP)*, Jun. 2020, pp. 283–286, doi: 10.1109/TENSYMP50017.2020.9230670.

[83] H. Gupta and M. Patel, ‘‘Method of text summarization using LSA and sentence based topic modelling with BERT,’’ in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 511–517, doi: 10.1109/ICAIS50930.2021.9395976.

[84] M. ShaukatTamboli and R. S. Prasad, ‘‘Authorship analysis and identification techniques: A review,’’ *Int. J. Comput. Appl.*, vol. 77, no. 16, pp. 11–15, Sep. 2013, doi: 10.5120/13566-1375.

[85] S. Hima Bindu Sri and S. R. Dutta, ‘‘A survey on automatic text summarization techniques,’’ *J. Phys., Conf. Ser.*, vol. 2040, no. 1, Oct. 2021, Art. no. 012044, doi: 10.1088/1742-6596/2040/1/012044.

[86] M. Potthast, A. Eiselt, L. A. B. Cedeño, B. Stein, and P. Rosso, ‘‘Overview of the 3rd international competition on plagiarism detection,’’ in *Proc. CEUR Workshop*, vol. 1177, 2011.

[87] V. Chandere, S. Satish, and R. Lakshminarayanan, ‘‘Online plagiarism detection tools in the digital age: A review,’’ *Ann. Rom. Soc. Cell Biol.*, vol. 25, no. 1, pp. 7110–7119, 2021. [Online]. Available: http://annalsofrscb.ro/index.php/journal/article/view/881

[88] H. Gruenthal, ‘‘An honest mistake: When an author omits a citation, is it plagiarism? One librarian has the courage to ask,’’ *Knowl. Quest*, vol. 37, no. 3, pp. 70–73, 2009.

[89] J. Beel, B. Gipp, S. Langer, and C. Breitinger, ‘‘Research-paper recommender systems: A literature survey,’’ *Int. J. Digit. Libraries*, vol. 17, no. 4, pp. 305–338, Nov. 2016, doi: 10.1007/S00799-015-0156-0/FIGURES/6.

[90] H. R. Garner, ‘‘Combating unethical publications with plagiarism detection services,’’ *Urol. Oncol., Seminars Original Invest.*, vol. 29, no. 1, pp. 95–99, Jan. 2011, doi: 10.1016/j.urolonc.2010.09.016.

[91] F. Hamborg, K. Donnay, and B. Gipp, ‘‘Automated identification of media bias in news articles: An interdisciplinary literature review,’’ *Int. J. Digit. Libraries*, vol. 20, no. 4, pp. 391–415, Dec. 2019, doi: 10.1007/S00799-018-0261-Y/FIGURES/2.

[92] S. V. Bruton, ‘‘Self-plagiarism and textual recycling: Legitimate forms of research misconduct,’’ *Accountability Res.*, vol. 21, no. 3, pp. 176–197, May 2014, doi: 10.1080/08989621.2014.848071.

[93] M. S. Sengupta, ‘‘Copyright infringement & plagiarism are they really two sides of a coin?’’ *CTBC’s IRJ*, vol. 2, no. 2, pp. 19–22, 2015.

[94] N. Lilian and J. Chukwuere, ‘‘The attitude of students towards plagiarism in online learning: A narrative literature review education view project meet African scholars (MAS) view project,’’ *Res. Publications*, vol. 18, pp. 14675–14688, Aug. 2020. [Online]. Available: https://www.researchgate.net/publication/343471863

[95] D. Pecorari and B. Petrić, "Plagiarism in second-language writing," *Lang. Teaching*, vol. 47, no. 3, pp. 269–302, Jul. 2014, doi: 10.1017/s0261444814000056.

[96] K. P. Gomes and S. N. Matos, "Detection of programming plagiarism in computing education: A systematic mapping study," *Cbie*, pp. 1633–1642, 2020, doi: 10.5753/cbie.sbie.2020.1633.

[97] H. Zha, "Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2002, pp. 113–120, doi: 10.1145/564396.564398.

[98] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 11265–11276.

[99] R. Ferreira, L. de Souza Cabral, F. Freitas, R. D. Lins, G. de França Silva, S. J. Simske, and L. Favaro, "A multi-document summarization system based on statistics and linguistic treatment," *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5780–5787, Oct. 2014, doi: 10.1016/j.eswa.2014.03.023.

[100] L. Almuqren and A. Cristea, "AraCust: A Saudi telecom tweets corpus for sentiment analysis," *PeerJ Comput. Sci.*, vol. 7, p. e510, May 2021, doi: 10.7717/peerj-cs.510.

[101] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguist. Conf. (ACL)*, vol. 1, 2018, pp. 328–339, doi: 10.18653/v1/p18-1031.

[102] K. Kurniawan and S. Louvan, "Indosum: A new benchmark dataset for Indonesian text summarization," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2018, pp. 215–220, doi: 10.1109/IALP.2018.8629109.

MUHAMMAD FARAZ MANZOOR is currently pursuing the Ph.D. degree with the University of Management and Technology, Lahore. He is a Senior Lecturer with Bahria University Lahore Campus. He has published many peer-reviewed international journals and conference papers. His research interests include natural language processing, machine learning, and deep learning.

MUHAMMAD SHOAIB FAROOQ was an Affiliate Member of George Mason University, USA. He is currently a Professor of artificial intelligence with the University of Management and Technology, Lahore. He possesses more than 28 years of teaching experience in the field of computer science. He has published many peer-reviewed international journals and conference papers. His research interests include the theory of programming languages, big data, the IoT, the Internet of Vehicles, machine learning, blockchain, and education.

MUHAMMAD HASEEB received the B.S. degree in computer science from GC University, Pakistan, in 2018, and the M.S. degree in computer science from the University of Management and Technology, Pakistan, in 2021. He worked as a Lecturer with the Department of Computer Science, NIIT, Lahore. His research interests include natural language processing and data science.

UZMA FAROOQ received the B.S. degree in computer science from the University of the Punjab, in 2005, the M.S. degree in computer science from the National University of Computer and Emerging Sciences, Pakistan, in 2007, and the Ph.D. degree from Universiti Teknologi Malaysia, Malaysia, in 2023. She is currently an Assistant Professor with the Department of Software Engineering, University of Management and Technology, Pakistan. She has published more than 20 papers in well reputed journals and conferences.

SOHAIL KHALID was with Emirates Airlines and Metlife. He is currently an Application Development Consultant with Saudi Aramco. He has 20 years of experience in software development. He is an Oracle Certified Professional Java as well as an Oracle Certified Expert in Java Web Services. His research interests include data analytics, natural language processing, and java programming.

ADNAN ABID (Senior Member, IEEE) received the Ph.D. degree from Politecnico di Milano, Italy, in 2012. He is currently a Professor of data science with the Faculty of Computing and Information Technology, University of the Punjab, Pakistan. He is a member of ACM. He has been associated with editorial boards of well-reputed journals in the area of computer science.