Academic Plagiarism Detection: A Systematic Literature Review

TOMÁŠ FOLTÝNEK, Department of Informatics, Mendel University in Brno, Czechia and University of Wuppertal, Germany
NORMAN MEUSCHKE and BELA GIPP, University of Wuppertal, Germany and University of Konstanz, Germany

This article summarizes the research on computational methods to detect academic plagiarism by systematically reviewing 239 research papers published between 2013 and 2018. To structure the presentation of the research contributions, we propose novel technically oriented typologies for plagiarism prevention and detection efforts, the forms of academic plagiarism, and computational plagiarism detection methods. We show that academic plagiarism detection is a highly active research field. Over the period we review, the field has seen major advances regarding the automated detection of strongly obfuscated and thus hard-to-identify forms of academic plagiarism. These improvements mainly originate from better semantic text analysis methods, the investigation of non-textual content features, and the application of machine learning. We identify a research gap in the lack of methodologically thorough performance evaluations of plagiarism detection systems. Concluding from our analysis, we see the integration of heterogeneous analysis methods for textual and non-textual content features using machine learning as the most promising area for future research contributions to improve the detection of academic plagiarism further.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → **Specialized information retrieval**; • **Computing methodologies** → **Natural language processing**; *Machine learning approaches*; • **Applied computing** → *Digital libraries and archives*;

Additional Key Words and Phrases: Plagiarism detection, literature review, text-matching software, semantic analysis, machine learning

ACM Reference format:
Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2019. Academic Plagiarism Detection: A Systematic Literature Review. *ACM Comput. Surv.* 52, 6, Article 112 (October 2019), 42 pages.
https://doi.org/10.1145/3345317

INTRODUCTION

Academic plagiarism is one of the severest forms of research misconduct (a "cardinal sin") [14] and has strong negative impacts on academia and the public. Plagiarized research papers impede

This work was supported by the EU ESF grant CZ.02.2.69/0.0/0.0/16_027/0007953 "MENDELU international development."
Authors' addresses: T. Foltýnek, Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czechia; email: tomas.foltynek@mendelu.cz; N. Meuschke and B. Gipp, Data & Knowledge Engineering Group, University of Wuppertal, School of Electrical, Information and Media Engineering, Rainer-Gruenter-Str. 21, D-42119 Wuppertal, Germany; emails: meuschke@uni-wuppertal.de, norman.meuschke@uni-konstanz.de, gipp@uni-wuppertal.de, bela.gipp@uni-konstanz.de.

0360-0300/2019/10-ART112
https://doi.org/10.1145/3345317

the scientific process, e.g., by distorting the mechanisms for tracing and correcting results. If researchers expand or revise earlier findings in subsequent research, then papers that plagiarized the original paper remain unaffected. Wrong findings can spread and affect later research or practical applications [90]. For example, in medicine or pharmacology, meta-studies are an important tool to assess the efficacy and safety of medical drugs and treatments. Plagiarized research papers can skew meta-studies and thus jeopardize patient safety [65].

Furthermore, academic plagiarism wastes resources. For example, Wager [261] quotes a journal editor stating that 10% of the papers submitted to the respective journal suffered from plagiarism of an unacceptable extent. In Germany, the ongoing crowdsourcing project VroniPlag[1] has investigated more than 200 cases of alleged academic plagiarism (as of July 2019). Even in the best case, i.e., if the plagiarism is discovered, reviewing and punishing plagiarized research papers and grant applications still causes a high effort for the reviewers, affected institutions, and funding agencies. The cases reported in VroniPlag showed that investigations into plagiarism allegations often require hundreds of work hours from affected institutions.

If plagiarism remains undiscovered, then the negative effects are even more severe. Plagiarists can unduly receive research funds and career advancements as funding agencies may award grants for plagiarized ideas or accept plagiarized research papers as the outcomes of research projects. The artificial inflation of publication and citation counts through plagiarism can further aggravate the problem. Studies showed that some plagiarized papers are cited at least as often as the original [23]. This phenomenon is problematic, since citation counts are widely used indicators of research performance, e.g., for funding or hiring decisions.

From an educational perspective, academic plagiarism is detrimental to competence acquisition and assessment. Practicing is crucial to human learning. If students receive credit for work done by others, then an important extrinsic motivation for acquiring knowledge and competences is reduced. Likewise, the assessment of competence is distorted, which again can result in undue career benefits for plagiarists.

The problem of academic plagiarism is not new but has been present for centuries. However, the rapid and continuous advancement of information technology (IT), which offers convenient and instant access to vast amounts of information, has made plagiarizing easier than ever. At the same time, IT also facilitated the detection of academic plagiarism. As we present in this article, hundreds of researchers address the automated detection of academic plagiarism and publish hundreds of research papers a year.

The high intensity and rapid pace of research on academic plagiarism detection make it difficult for researchers to get an overview of the field. Published literature reviews alleviate the problem by summarizing previous research, critically examining contributions, explaining results, and clarifying alternative views [212, 40]. Literature reviews are particularly helpful for young researchers and researchers who newly enter a field. Often, these two groups of researchers contribute new ideas that keep a field alive and advance the state of the art.

In 2013, we provided a first descriptive review of the state of the art in academic plagiarism detection [160]. Given the rapid development of the field, we see the need for a follow-up study to summarize the research since 2013. Therefore, this article provides a systematic qualitative literature review [187] that critically evaluates the capabilities of computational methods to detect plagiarism in academic documents and identifies current research trends and research gaps.

The literature review at hand answers the following research questions:

1. What are the major developments in the research on computational methods for plagiarism detection in academic documents since our last literature review in 2013?

[1] http://de.vroniplag.wikia.com.

   a. Did researchers propose conceptually new approaches for this task?
   b. Which improvements to existing detection methods have been reported?
2. Which research gaps and trends for future research are observable in the literature?

To answer these questions, we organize the remainder of this article as follows. The section *Methodology* describes our procedure and criteria for data collection. The following section, *Related Literature Reviews*, summarizes the contributions of our compared to topically related reviews published since 2013. The section *Overview of the Research Field* describes the major research areas in the field of academic plagiarism detection. The section *Definition and Typology of Plagiarism* introduces our definition and a three-layered model for addressing plagiarism (methods, systems, and policies). The section *Review of Plagiarism Typologies* synthesizes the classifications of plagiarism found in the literature into a technically oriented typology suitable for our review. The section *Plagiarism Detection Methods* is the core of this article. For each class of computational plagiarism detection methods, the section provides a description and an overview of research papers that employ the method in question. The section *Plagiarism Detection Systems* discusses the application of detection methods in plagiarism detection systems. The *Discussion* section summarizes the advances in plagiarism detection research and outlines open research questions.

METHODOLOGY

To collect the research papers included in our review, we performed a keyword-based automated search [212] using Google Scholar and Web of Science. We limited the search period to 2013 until 2018 (including). However, papers that introduced a novel concept or approach often predate 2013. To ensure that our survey covers all relevant primary literature, we included such seminal papers regardless of their publication date.

Google Scholar indexes major computer science literature databases, including IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and TandFonline, as well as grey literature. Fagan [68] provides an extensive list of "*recent studies [that] repeatedly find that Google Scholar's coverage meets or exceeds that of other search tools, no matter what is identified by target samples, including journals, articles, and citations*" [68]. Therefore, we consider Google Scholar as a meta-database that meets the search criteria recommended in the guidelines for conducting systematic literature reviews [40, 137]. Using Google Scholar also addresses the *"lack of conformity, especially in terms of searching facilities, across commonly used digital libraries,"* which Brereton et al. [40] identified as a hindrance to systematic literature reviews in computer science.

Criticism of using Google Scholar for literature research includes that the system's relevance ranking assigns too much importance to citation count [68], i.e., the number of citations a paper receives. Moreover, Google Scholar covers predatory journals [31]. Most guidelines for systematic reviews, therefore, recommend using additional search tools despite the comprehensive coverage of Google Scholar [68]. Following this recommendation, we additionally queried Web of Science. Since we seek to cover the most influential papers on academic plagiarism detection, we consider a relevance ranking based on citation counts as an advantage rather than a disadvantage. Hence, we used the relevance ranking of Google Scholar and ranked search results from Web of Science by citation count. We excluded all papers (11) that appeared in venues mentioned in *Beall's List of Predatory Journals and Publishers*.[2]

Our procedure for paper collection consisted of the five phases described hereafter. We reviewed the first 50 search results when using Google Scholar and the first 150 search results when using Web of Science.

[2] https://beallslist.weebly.com/standalone-journals.html.

In the first phase, we sought to include existing literature reviews on plagiarism detection for academic documents. Therefore, we queried Google Scholar using the following keywords: *plagiarism detection literature review, similarity detection literature review, plagiarism detection state of art, similarity detection state of art, plagiarism detection survey, similarity detection survey.*

In the second phase, we added topically related papers using the following rather general keywords: *plagiarism, plagiarism detection, similarity detection, extrinsic plagiarism detection, external plagiarism detection, intrinsic plagiarism detection, internal plagiarism detection.*

After reviewing the papers retrieved in the first and second phases, we defined the structure of our review and adjusted the scope of our data collection as follows:

1. We focused our search on plagiarism detection for text documents and hence excluded papers addressing other tasks, such as plagiarism detection for source code or images. We also excluded papers focusing on corpora development.
2. We excluded papers addressing policy and educational issues related to plagiarism detection to sharpen the focus of our review on computational detection methods.

Having made these adjustments to our search strategy, we started the third phase of the data collection. We queried Google Scholar with the following keywords related to specific sub-topics of plagiarism detection, which we had identified as important during the first and second phases: *semantic analysis plagiarism detection, machine-learning plagiarism detection.*

In the fourth phase, we sought to prevent selection bias from exclusively using Google Scholar by querying Web of Science using the keyword *plagiarism detection.*

In the fifth phase, we added to our dataset papers from the search period that are topically related to papers we had already collected. To do so, we included relevant references of collected papers and papers that publishers' systems recommended as related to papers in our collection. Following this procedure, we included notebook papers of the annual PAN and SemEval workshops. To ensure the significance of research contributions, we excluded papers that were not referenced in the official overview papers of the PAN and SemEval workshops or reported results below the baseline provided by the workshop organizers. For the same reason, we excluded papers that do not report experimental evaluation results.

To ensure the consistency of paper processing, the first author read all papers in the final dataset and recorded the paper's key content in a mind map. All authors continuously reviewed, discussed, and updated the mind map. Additionally, we maintained a spreadsheet to record the key features of each paper (task, methods, improvements, dataset, results, etc.).

Table 1 and Table 2 list the numbers of papers retrieved and processed in each phase of the data collection.

Methodological Risks

The main risks for systematic literature reviews are incompleteness of the collected data and deficiencies in the selection, structure, and presentation of the content.

We addressed the risk of data incompleteness mainly by using two of the most comprehensive databases for academic literature—Google Scholar and Web of Science. To achieve the best possible coverage, we queried the two databases with keywords that we gradually refined in a multi-stage process, in which the results of each phase informed the next phase. By including all relevant references of papers that our keyword-based search had retrieved, we leveraged the knowledge of domain experts, i.e., the authors of research papers and literature reviews on the topic, to retrieve additional papers. We also included the content-based recommendations provided by the digital library systems of major publishers, such as Elsevier and ACM. We are confident that this

Table 1. Numbers of Papers in Each Phase of the Data Collection

Phase	No. of Papers	Excluded (source code, policy, corpora)	Newly Added	Already Included	Collection Size
1) Google Scholar: reviews	66	28	38	–	38
2) Google Scholar: related papers	143	54	89	23	104
3) Google Scholar: sub-topics	–	–	49	42	111
4) Web of Science	134	82	52	35	128
5) Processing stage	126	–	126	–	254

Table 2. Numbers of Papers by Category

Papers identified by keyword-based automated search	128
Papers collected through references and automated recommendations	126
Inaccessible papers	3
Excluded papers	12
Number of papers reviewed	**239**
- Reviews and general papers	35
- Papers containing experiments (included in overview tables)	204
– Extrinsic PD	136
– Intrinsic PD	67
– Both extrinsic and intrinsic PD	1

multi-faceted and multi-stage approach to data collection yielded a set of papers that comprehensively reflects the state of the art in detecting academic plagiarism.

To mitigate the risk of subjectivity regarding the selection and presentation of content, we adhered to best practice guidelines for conducting systematic reviews and investigated the taxonomies and structure put forward in related reviews. We present the insights of the latter investigation in the following section.

RELATED LITERATURE REVIEWS

Table 3 lists related literature reviews in chronological order and categorized according to (i) the plagiarism detection (PD) tasks the review covers (PD for text documents, PD for source code, other PD tasks), (ii) whether the review includes descriptions or evaluations of productive plagiarism detection systems, and (iii) whether the review addresses policy issues related to plagiarism and academic integrity. All reviews are "narrative" according to the typology of Pare et al. [187]. Two of the reviews (References [61] and [48]) cover articles that appeared at venues included in *Beall's List of Predatory Journals and Publishers*.

Our previous review article [160] surveyed the state of the art in detecting academic plagiarism, presented plagiarism detection systems, and summarized evaluations of their detection effectiveness. We outlined the limitations of text-based plagiarism detection methods and suggested that future research should focus on semantic analysis approaches that also include non-textual document features, such as academic citations.

Table 3. Related Literature Reviews on Plagiarism Detection

	Plagiarism Detection Tasks			Systems	Policies
Paper	**Text**	**Source Code**	**Other**		
Meuschke and Gipp [160]	YES	NO	NO	YES	NO
Chong [47]	YES	NO	NO	NO	NO
Eisa et al. [61]	YES	NO	YES	NO	NO
Agarwal and Sharma [8]	YES	YES	NO	YES	NO
Chowdhury et al. [48]	YES	YES	NO	YES	NO
Kanjirangat and Gupta [251]	YES	YES	NO	YES	NO
Velasquez et al. [256]	YES	NO	NO	YES	YES
Hourrane and Benlahmar [114]	YES	NO	NO	NO	NO

The main contribution of Chong [47] is an extensive experimental evaluation of text preprocessing methods as well as shallow and deep NLP techniques. However, the paper also provides a sizable state-of-the-art review of plagiarism detection methods for text documents.

Eisa et al. [61] defined a clear methodology and meticulously followed it but did not include a temporal dimension. Their well-written review provides comprehensive descriptions and a useful taxonomy of features and methods for plagiarism detection. The authors concluded that future research should consider non-textual document features, such as equations, figures, and tables.

Agarwal and Sharma [8] focused on source code PD but also gave a basic overview of plagiarism detection methods for text documents. Technologically, source code PD and PD for text are closely related, and many plagiarism detection methods for text can also be applied for source code PD [57].

Chowdhury et al. [48] provided a comprehensive list of available plagiarism detection systems.

Kanjirangat and Gupta [251] summarized plagiarism detection methods for text documents that participated in the PAN competitions and compared four plagiarism detection systems.

Velasquez et al. [256] proposed a new plagiarism detection system but also provided an extensive literature review that includes a typology of plagiarism and an overview of six plagiarism detection systems.

Hourrane and Benlahmar [114] described individual research papers in detail but did not provide an abstraction of the presented detection methods.

The literature review at hand extends and improves the reviews outlined in Table 3 as follows:

1. We include significantly more papers than other reviews.
2. Our literature survey is the first that analyses research contributions during a specific period to provide insights on the most recent research trends.
3. Our review is the first that adheres to the guidelines for conducting systematic literature surveys.
4. We introduce a three-layered conceptual model to describe and analyze the phenomenon of academic plagiarism comprehensively.

OVERVIEW OF THE RESEARCH FIELD

The papers we retrieved during our research fall into three broad categories: **plagiarism detection methods, plagiarism detection systems**, and **plagiarism policies**. Ordering these categories by the level of abstraction at which they address the problem of academic plagiarism yields the three-layered model shown in Figure 1. We propose this model to structure and systematically analyze the large and heterogeneous body of literature on academic plagiarism.

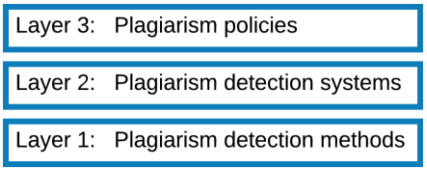

Fig. 1. Three-layered model for addressing academic plagiarism.

Layer 1: Plagiarism detection methods subsumes research that addresses the automated identification of potential plagiarism instances. Papers falling into this layer typically present methods that analyze textual similarity at the lexical, syntactic, and semantic levels, as well as similarity of non-textual content elements, such as citations, figures, tables, and mathematical formulae. To this layer, we also assign papers that address the evaluation of plagiarism detection methods, e.g., by providing test collections and reporting on performance comparisons. The research contributions in Layer 1 are the focus of this survey.

Layer 2: Plagiarism detection systems encompasses applied research papers that address production-ready plagiarism detection systems, as opposed to the research prototypes that are typically presented in papers assigned to Layer 1. Production-ready systems implement the detection methods included in Layer 1, visually present detection results to the users and should be able to identify duly quoted text. Turnitin LLC is the market leader for plagiarism detection services. The company's plagiarism detection system Turnitin is most frequently cited in papers included in Layer 2 [116, 191, 256].

Layer 3: Plagiarism policies subsumes papers that research the prevention, detection, prosecution, and punishment of plagiarism at educational institutions. Typical papers in Layer 3 investigate students' and teachers' attitudes toward plagiarism (e.g., Reference [75]), analyze the prevalence of plagiarism at institutions (e.g., Reference [50]), or discuss the impact of institutional policies (e.g., Reference [183]).

The three layers of the model are interdependent and essential to analyze the phenomenon of academic plagiarism comprehensively. Plagiarism detection systems (Layer 2) depend on reliable detection methods (Layer 1), which in turn would be of little practical value without production-ready systems that employ them. Using plagiarism detection systems in practice would be futile without the presence of a policy framework (Layer 3) that governs the investigation, documentation, prosecution, and punishment of plagiarism. The insights derived from analyzing the use of plagiarism detection systems in practice (Layer 3) also inform the research and development efforts for improving plagiarism detection methods (Layer 1) and plagiarism detection systems (Layer 2).

Continued research in all three layers is necessary to keep pace with the behavior changes that are a typical reaction of plagiarists when being confronted with an increased risk of discovery due to better detection technology and stricter policies. For example, improved plagiarism detection capabilities led to a rise in contract cheating, i.e., paying ghostwriters to produce original works that the cheaters submit as their own [177]. Many researchers agree that counteracting these developments requires approaches that integrate plagiarism detection technology with plagiarism policies.

Originally, we intended to survey the research in all three layers. However, the extent of the research fields is too large to cover all of them in one survey comprehensively. Therefore, the current article surveys plagiarism detection methods and systems. A future survey will cover the research on plagiarism policies.

DEFINITION AND TYPOLOGY OF PLAGIARISM

In accordance with Fishman, we define academic plagiarism as the use of ideas, content, or structures without appropriately acknowledging the source to benefit in a setting where originality is

expected [279]. We used a nearly identical definition in our previous survey [160], because it describes the full breadth of the phenomenon. The definition includes all forms of intellectual contributions in academic documents regardless of their presentation, e.g., text, figures, tables, and mathematical formulae, and their origin. Other definitions of academic plagiarism often include the notion of theft (e.g., References [13, 38, 116, 146, 188, 274, 252]), i.e., require intent and limit the scope to reusing the content of others. Our definition also includes self-plagiarism, unintentional plagiarism, and plagiarism with the consent of the original author.

Review of Plagiarism Typologies

Aside from a definition, a typology helps to structure the research and facilitates communication on a phenomenon [29, 261]. Researchers proposed a variety of typologies for academic plagiarism. Walker [263] coined a typology from a plagiarist's point of view, which is still recognized by contemporary literature [51]. Walker's typology distinguishes between:

1. Sham paraphrasing (*presenting copied text as a paraphrase by leaving out quotations*)
2. Illicit paraphrasing
3. Other plagiarism (*plagiarizing with the original author's consent*)
4. Verbatim copying (*without reference*)
5. Recycling (*self-plagiarism*)
6. Ghostwriting
7. Purloining (*copying another student's assignment without consent*)

(Notes in parentheses denote explanations we added for clarity.)

All typologies we encountered in our research categorize verbatim copying as one form of academic plagiarism. Alfikri and Ayu Purwarianti [13] additionally distinguished as separate forms of academic plagiarism the partial copying of smaller text segments, two forms of paraphrasing that differ regarding whether the sentence structure changes and translations. Velasquez et al. [256] distinguished verbatim copying and technical disguise, combined paraphrasing and translation into one form, and categorized the deliberate misuse of references as a separate form. Weber-Wulff [265] and Chowdhury and Bhattacharyya [48] likewise categorized referencing errors as a form of plagiarism. Many authors agreed on classifying idea plagiarism as a separate form of plagiarism [47, 48, 114, 179, 252]. Mozgovoy et al. [173] presented a typology that consolidates other classifications into five forms of academic plagiarism:

1. Verbatim copying
2. Hiding plagiarism instances by paraphrasing
3. Technical tricks exploiting weaknesses of current plagiarism detection systems
4. Deliberately inaccurate use of references
5. Tough plagiarism

"Tough plagiarism" subsumes the forms of plagiarism that are difficult to detect for both humans and computers, like idea plagiarism, structural plagiarism, and cross-language plagiarism [173].

The typology of Eisa et al. [61], which originated from a typology by Alzahrani et al. [21], distinguishes only two forms of plagiarism: *literal plagiarism* and *intelligent plagiarism*. Literal plagiarism encompasses near copies and modified copies, whereas intelligent plagiarism includes paraphrasing, summarization, translation, and idea plagiarism.

Our Typology of Plagiarism

Since we focus on reviewing plagiarism detection technology, we exclusively consider technical properties to derive a typology of academic plagiarism forms. From a technical perspective, several

distinctions that are important from a policy perspective are irrelevant or at least less important. Technically irrelevant properties of plagiarism instances are whether:

- the original author permitted to reuse content;
- the suspicious document and its potential source have the same author(s), i.e., whether similarities in the documents' content may constitute self-plagiarism.

Properties of minor technical importance are:

- how much of the content represents potential plagiarism;
- whether a plagiarist uses one or multiple sources. Detecting compilation plagiarism (also referred to as shake-and-paste, patch-writing, remix, mosaic or mash-up) is impossible at the document level but requires an analysis on the level of paragraphs or sentences.

Both properties are of little technical importance, since similar methods are employed regardless of the extent of plagiarism and whether it may originate from one or multiple source documents.

Our typology of academic plagiarism derives from the generally accepted layers of natural language: lexis, syntax, and semantics. Ultimately, the goal of language is expressing ideas [96]. Therefore, we extend the classic three-layered language model to four layers and categorize plagiarism forms according to the language layer they affect. We order the resulting plagiarism forms increasingly by their level of obfuscation:

1. Characters-preserving plagiarism
   - Literal plagiarism (copy and paste)
   - Possibly with mentioning the source
2. Syntax-preserving plagiarism
   - Technical disguise
   - Synonym substitution
3. Semantics-preserving plagiarism
   - Translation
   - Paraphrase (mosaic, clause quilts)
4. Idea-preserving plagiarism
   - Structural plagiarism
   - Using concepts and ideas only
5. Ghostwriting

Characters-preserving plagiarism includes, aside from verbatim copying, plagiarism forms in which sources are mentioned, like "pawn sacrifice" and "cut and slide" [265]. *Syntax-preserving plagiarism* often results from employing simple substitution techniques, e.g., using regular expressions. Basic synonym substitution approaches operate in the same way; however, employing more sophisticated substitution methods has become typical. *Semantics-preserving plagiarism* refers to sophisticated forms of obfuscation that involve changing both the words and the sentence structure but preserve the meaning of passages. In agreement with Velasquez et al. [256], we consider translation plagiarism as a semantics-preserving form of plagiarism, since a translation can be seen as the ultimate paraphrase. In the section devoted to semantics-based plagiarism detection methods, we will also show a significant overlap in the methods for paraphrase detection and cross-language plagiarism detection. *Idea-preserving plagiarism* (also referred to as template plagiarism or boilerplate plagiarism) includes cases in which plagiarists use the concept or structure of a source and describe it entirely in their own words. This form of plagiarism is difficult to identify and even harder to prove. *Ghostwriting* [47, 114] describes the hiring of a third party to write genuine text [50, 263]. It is the only form of plagiarism that is undetectable by comparing

a suspicious document to a likely source. Currently, the only technical option for discovering potential ghostwriting is to compare stylometric features of a possibly ghost-written document with documents certainly written by the alleged author.

PLAGIARISM DETECTION APPROACHES

Conceptually, the task of detecting plagiarism in academic documents consists of locating the parts of a document that exhibit indicators of potential plagiarism and subsequently substantiating the suspicion through more in-depth analysis steps [218]. From a technical perspective, the literature distinguishes the following two general approaches to plagiarism detection.

The ***extrinsic plagiarism detection*** approach compares suspicious documents to a collection of documents assumed to be genuine (reference collection) and retrieves all documents that exhibit similarities above a threshold as potential sources [252, 235].

The ***intrinsic plagiarism detection*** approach exclusively analyzes the input document, i.e., does not perform comparisons to documents in a reference collection. Intrinsic detection methods employ a process known as *stylometry* to examine linguistic features of a text [90]. The goal is to identify changes in writing style, which the approach considers as indicators for potential plagiarism [277]. Passages with linguistic differences can become the input for an extrinsic plagiarism analysis or be presented to human reviewers. Hereafter, we describe the extrinsic and intrinsic approaches to plagiarism detection in more detail.

Extrinsic Plagiarism Detection

The reference collection to which extrinsic plagiarism detection approaches compare the suspicious document is typically very large, e.g., a significant subset of the Internet for production-ready plagiarism detection systems. Therefore, pairwise comparisons of the input document to all documents in the reference collection are often computationally infeasible. To address this challenge, most extrinsic plagiarism detection approaches consist of two stages: *candidate retrieval* (also called source retrieval) and *detailed analysis* (also referred to as text alignment) [197]. The candidate retrieval stage efficiently limits the collection to a subset of potential source documents. The detailed analysis stage then performs elaborate pairwise document comparisons to identify parts of the source documents that are similar to parts of the suspicious document.

Candidate Retrieval. Given a suspicious input document and a querying tool, e.g., a search engine or database interface, the task in the candidate retrieval stage is to retrieve from the reference collection all documents that share content with the input document [198]. Many plagiarism detection systems use the APIs of Web search engines instead of maintaining own reference collections and querying tools.

Recall is the most important performance metric for the candidate retrieval stage of the extrinsic plagiarism detection process, since the subsequent detailed analysis cannot identify source documents missed in the first stage [105]. The number of queries issued is another typical metric to quantify the performance in the candidate retrieval stage. Keeping the number of queries low is particularly important if the candidate retrieval approach involves Web search engines, since such engines typically charge for issuing queries.

Detailed Analysis. The set of documents retrieved in the candidate retrieval stage is the input to the detailed analysis stage. Formally, the task in the detailed analysis stage is defined as follows. Let d_q be a suspicious document. Let $D = \{d_s\} \mid s = 1 \ldots n$ be a set of potential source documents. Determine whether a fragment $s_q \in d_q$ is similar to a fragment $s \in d_s$ ($d_s \in D$) and identify all such pairs of fragments (s_q, s) [202]. Eventually, an expert should determine whether the identified

pairs (s_q, s) constitute legitimate content re-use, plagiarism, or false positives [29]. The detailed analysis typically consists of three steps [197]:

1. *Seeding*: Finding parts of the content in the input document (the seed) within a document of the reference collection
2. *Extension*: Extending each seed as far as possible to find the complete passage that may have been reused
3. *Filtering*: Excluding fragments that do not meet predefined criteria (e.g., that are too short), and handling of overlapping passages

The most common strategy for the extension step is the so-called rule-based approach. The approach merges seeds if they occur next to each other in both the suspicious and the source document and if the size of the gap between the passages is below a threshold [198].

Paraphrase Identification is often a separate step within the detailed analysis stages of extrinsic plagiarism detection methods but also a research field on its own. The task in paraphrase identification is determining semantically equivalent sentences in a set of sentences [71]. SemEval is a well-known conference series that addresses paraphrase identification for tweets [9, 222]. Identifying semantically equivalent tweets is more difficult than identifying semantically equivalent sentences in academic documents due to out-of-vocabulary words, abbreviations, and slang terms that are frequent in tweets [24]. Al-Samadi et al. [9] provided a thorough review of the research on paraphrase identification.

Intrinsic Plagiarism Detection

The concept of intrinsic plagiarism detection was introduced by Meyer zu Eissen and Stein [277]. Whereas extrinsic plagiarism detection methods search for similarities across documents, intrinsic plagiarism detection methods search for dissimilarities within a document. A crucial presumption of the intrinsic approach is that authors have different writing styles that allow identifying the authors. Juola provides a comprehensive overview of stylometric methods to analyze and quantify writing style [127].

Intrinsic plagiarism detection consists of two tasks [200, 233]:

1. *Style breach detection*: Delineating passages with different writing styles
2. *Author identification*: Identifying the author of documents or passages

Author identification furthermore subsumes two specialized tasks:

1. *Author clustering*: Grouping documents or passages by authorship
2. *Author verification*: Deciding whether an input document was authored by the same person as a set of sample documents

Style Breach Detection. Given a suspicious document, the goal of style-breach detection is identifying passages that exhibit different stylometric characteristics [233]. Most of the algorithms for style breach detection follow a three-step process [214]:

1. *Text segmentation* based on paragraphs, (overlapping) sentences, character or word n-grams
2. *Feature space mapping*, i.e., computing stylometric measures for segments
3. *Clustering* segments according to observed critical values

Author Clustering typically follows the style breach detection stage and employs pairwise comparisons of passages identified in the previous stage to group them by author [247]. For each pair of passages, a similarity measure is computed that considers the results of the feature space

mapping in the style-breach detection stage. Formally, for a given set of documents or passages D, the task is to find the decomposition of this set $D_1, D_2, \ldots Dn$, such that:

1. $D = \mathrm{U}_{i=1}^{n} D_i$
2. $D_i \cap D_j = \emptyset$ for each $i \neq j$
3. All documents of the same class have the same author;
   For each pair of documents from different classes, the authors are different.

Author Verification is typically defined as the prediction of whether two pieces of text were written by the same person. In practice, author verification is a one-class classification problem [234] that assumes all documents in a set have the same author. By comparing the writing style at the document level, outliers can be detected that may represent plagiarized documents. This method can reveal ghostwriting [127], unless the same ghost-writer authored all documents in the set.

Author Identification (also referred to as author classification), takes multiple document sets as input. Each set of documents must have been written verifiably by a single author. The task is assigning documents with unclear authorship to the stylistically most similar document set. Each authorship identification problem, for which the set of candidate authors is known, is easily transformable into multiple authorship verification problems [128]. An open-set variant of the author identification problem allows for a suspicious document with an author that is not included in any of the input sets [234].

Several other stylometry-based tasks, e.g., author profiling, exist. However, we limit the descriptions in the next section to methods whose main application is plagiarism detection. We recommend readers interested in related tasks to refer to the overview paper of PAN'17 [200].

PLAGIARISM DETECTION METHODS

We categorize plagiarism detection methods and structure their description according to our typology of plagiarism. *Lexical detection methods* exclusively consider the characters in a document. *Syntax-based detection methods* consider the sentence structure, i.e., the parts of speech and their relationships. *Semantics-based detection methods* compare the meaning of sentences, paragraphs, or documents. *Idea-based detection methods* go beyond the analysis of text in a document by considering non-textual content elements like citations, images, and mathematical content. Before presenting details on each class of detection methods, we describe preprocessing strategies that are relevant for all classes of detection methods.

Preprocessing

The initial preprocessing steps applied as part of plagiarism detection methods typically include document format conversions and information extraction. Before 2013, researchers described the extraction of text from binary document formats like PDF and DOC as well as from structured document formats like HTML and DOCX in more details than in more recent years (e.g., Reference [49]). Most research papers on text-based plagiarism detection methods we review in this article do not describe any format conversion or text extraction procedures. We attribute this development to the technical maturity of text extraction approaches. For plagiarism detection approaches that analyze non-textual content elements, e.g., academic citations and references [90, 91, 161, 191], images [162], and mathematical content [163, 165], document format conversion, and information extraction still present significant challenges.

Specific preprocessing operations heavily depend on the chosen approach. The aim is to remove noise while keeping the information required for the analysis. For text-based detection methods, typical preprocessing steps include lowercasing, punctuation removal, tokenization, segmentation, number removal or number replacement, named entity recognition, stop words removal,

stemming or lemmatization, Part of Speech (PoS) tagging, and synset extension. Approaches employing synset extension typically employ thesauri like WordNet [69] to assign the identifier of the class of synonymous words to which a word in the text belongs. The synonymous words can then be considered for similarity calculation. Detection methods operating on the lexical level usually perform chunking as a preprocessing step. Chunking groups text elements into sets of given lengths, e.g., word n-grams, line chunks, or phrasal constituents in a sentence [47].

Some detection approaches, especially in intrinsic plagiarism detection, limit preprocessing to a minimum to not loose potentially useful information [9, 67]. For example, intrinsic detection methods typically do not remove punctuation.

All preprocessing steps we described represent standard procedures in Natural Language Processing (NLP), hence well-established, publicly available software libraries support these steps. The research papers we reviewed predominantly used the multilingual and multifunctional text processing pipelines Natural Language Toolkit Kit (Python) or Stanford CoreNLP library (Java). Commonly applied syntax analysis tools include Penn Treebank,[3] Citar,[4] TreeTagger,[5] and Stanford parser.[6] Several papers present resources for Arabic [33, 34, 227] and Urdu [54] language processing.

Lexical Detection Methods

Lexical detection methods exclusively consider the characters in a text for similarity computation. The methods are best suited for identifying copy-and-paste plagiarism that exhibits little to no obfuscation. To detect obfuscated plagiarism, the lexical detection methods must be combined with more sophisticated NLP approaches [9, 67]. Lexical detection methods are also well-suited to identify homoglyph substitutions, which are a common form of technical disguise. The only paper in our collection that addressed the identification of technically disguised plagiarism is Reference [19]. The authors used a list of confusable Unicode characters and applied approximate word n-gram matching using the normalized Hamming distance.

Lexical detection approaches typically fall into one of the three categories we describe in the following: *n-gram comparisons, vector space models,* and *querying search engines.*

N-gram Comparisons. Comparing n-grams refers to determining the similarity of sequences of n consecutive entities, which are typically characters or words and less frequently phrases or sentences. n-gram comparisons are widely applied for candidate retrieval or the seeding phase of the detailed analysis stage in extrinsic monolingual and cross-language detection approaches as well as in intrinsic detection.

Approaches using n-gram comparisons first split a document into (possibly overlapping) n-grams, which they use to create a set-based representation of the document or passage ("fingerprint"). To enable efficient retrieval, most approaches store fingerprints in index data structures. To speed up the comparison of individual fingerprints, some approaches hash or compress the n-grams that form the fingerprints. Hashing or compression reduces the lengths of the strings under comparison and allows performing computationally more efficient numerical comparisons. However, hashing introduces the risk of false positives due to hash collisions. Therefore, hashed or compressed fingerprinting is more commonly applied for the candidate retrieval stage, in which achieving high recall is more important than achieving high precision.

[3] https://www.ldc.upenn.edu/.
[4] http://github.com/danieldk/citar.
[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
[6] http://nlp.stanford.edu/software/lex-parser.shtml.

Table 4. Word n-gram Detection Methods

Approach	Task	Method variation	Papers
Extrinsic	Document-level detection	Stop words removed	[20, 102, 126, 156, 195, 228, 256]
		Stop word n-grams	[101]
	Candidate retrieval	Stop words removed	[58]
		All word n-grams and stop word n-grams	[239]
	Detailed analysis	All word n-grams	[19, 185, 225]
		Stop words removed	[97, 100]
		All word n-grams, stop word n-grams, and named entity n-grams	[226]
		Numerous n-gram variations	[3, 186]
		Context n-grams	[210, 244]
	Paraphrase identification	All word n-grams	[113, 243]
		Combination with ESA	[260]
	CLPD	Stop words removed	[60]
Intrinsic	Author identification	Overlap in LZW dictionary	[45]
	Author verification	Word n-grams	[81, 103, 122, 135, 159, 170, 219]
		Stop word n-grams	[135, 159, 170]

Fingerprinting is the most popular method for assessing local lexical similarity [104]. However, recent research has focused increasingly on detecting obfuscated plagiarism. Thus n-gram fingerprinting is often restricted to the preprocessing stage [20] or used as a feature for machine learning [7]. Character n-gram comparisons can be applied to cross-language plagiarism detection (CLPD) if the languages in question exhibit a high lexical similarity, e.g., English and Spanish [79].

Table 4 presents papers employing word n-grams; Table 5 lists papers using character n-grams, and Table 6 shows papers that employ hashing or compression for n-gram fingerprinting.

Vector Space Models (VSM) are a classic retrieval approach that represents texts as high-dimensional vectors [249]. In plagiarism detection, words or word n-grams typically form the dimensions of the vector space and the components of a vector undergo term frequency–inverse document frequency (tf-idf) weighting [249]. Idf values are either derived from the suspicious document or the corpus [205, 238]. The similarity of vector representations—typically quantified using the cosine measure, i.e., the angle between the vectors—is used as a proxy for the similarity of the documents the vectors represent.

Most approaches employ predefined similarity thresholds to retrieve documents or passages for subsequent processing. Kanjirangat and Gupta [249] and Ravi et al. [208] follow a different approach. They divide the set of source documents into K clusters by first selecting K centroids and then assigning each document to the group whose centroid is most similar. The suspicious document is used as one of the centroids and the corresponding cluster is passed on to the subsequent processing stages.

VSM remain popular and well-performing approaches not only for detecting copy-and-paste plagiarism but also for identifying obfuscated plagiarism as part of a semantic analysis. VSM are also frequently applied in intrinsic plagiarism detection. A typical approach is to represent sentences as vectors of stylometric features to find outliers or to group stylistically similar sentences.

Table 5. Character n-gram (CNG) Detection Methods

Approach	Task	Method variation	Papers
Extrinsic	Document-level detection	Pure character n-grams	[178, 267]
		Overlap in LZW dictionary	[231]
		Machine learning	[38]
		Combined with Bloom filters	[86]
	Detailed analysis	Hashed character n-grams	[18]
	Paraphrase identification	Feature for machine learning	[243]
	Cross-language PD	Cross-language CNG	[28, 72, 74, 78]
Intrinsic	Style-breach detection	CNG as stylometric features	[35, 148],
	Author identification	Bit n-grams	[190]
	Author verification	CNG as stylometric features	[30, 44, 52, 81, 108], [109, 110, 121, 122, 123, 135, 146, 150, 157, 170, 219, 221]
	Author clustering	CNG as stylometric features	[10, 84, 98, 141, 220]

Table 6. Detection Methods Employing Compression or Hashing

Task	Method	Papers
Document-level detection	Hashing	[86, 83, 228, 267]
Candidate retrieval	Hashing	[20, 104, 175, 180]
Detailed analysis	Hashing	[18, 181, 185, 186]
Document-level detection	Compression	[231]
Author identification	Compression	[37, 45, 107, 112]

Table 7 presents papers that employ VSM for extrinsic plagiarism detection; Table 8 lists papers using VSM for intrinsic plagiarism detection.

Querying Web Search Engines. Many detection methods employ Web search engines for candidate retrieval, i.e., for finding potential source documents in the initial stage of the detection process. The strategy for selecting the query terms from the suspicious document is crucial for the success of this approach. Table 9 gives an overview of the strategies for query term selection employed by papers in our collection.

Intrinsic detection approaches can employ Web Search engines to realize the G*eneral Impostors Method*. This method transforms the one-class verification problem regarding an author's writing style into a two-class classification problem. The method extracts keywords from the suspicious document to retrieve a set of topically related documents from external sources, the so-called "impostors." The method then quantifies the "average" writing style observable in impostor documents, i.e., the distribution of stylistic features to be expected. Subsequently, the method compares the stylometric features of passages from the suspicious document to the features of the "average" writing style in impostor documents. This way, the method distinguishes the stylistic features that are characteristic of an author from the features that are specific to the topic [135]. Koppel and Winter present the method in detail [146]. Detection approaches implementing the general impostors method achieved excellent results in the PAN competitions, e.g., winning the competition in 2013 and 2014 [128, 232]. Table 10 presents papers using this method.

Table 7. Extrinsic Detection Methods Employing Vector Space Models

Task	Unit	Extension of VSM	Papers
Document-level detection	sentence	Combination of similarity metrics	[250]
Document-level detection	sentence	VSM as a bitmap; compressed for comparison	[223]
Document-level detection	sentence	Machine learning to set similarity thresholds	[88]
Document-level detection	word	Synonym replacement	[85]
Document-level detection	word, sentence	Fuzzy set of WordNet synonyms	[178]
Candidate retrieval	word	Vectors of word N-grams	[60, 125, 253],
Candidate retrieval	word	K-means clustering of vectors to find documents most similar to the input doc.	[208, 249]
Candidate retrieval	word	Z-order mapping of multidimensional vectors to scalar and subsequent filtering	[12]
Candidate retrieval	word	Topic-based segmentation; Re-ranking of results based on the proximity of terms	[58]
Detailed analysis	sentence	Pure VSM	[59, 145, 186, 252]
Detailed analysis	sentence	Adaptive adjustment of parameters to detect the type of obfuscation	[216, 217]
Detailed analysis	sentence	Hybrid similarity (Cosine+ Jaccard)	[264]
Detailed analysis	word	Pure VSM	[156]
Paraphrase identification	sentence	Semantic role annotation	[71]

Table 8. Intrinsic Detection Methods Employing Vector Space Models

Task	Unit	Extension of VSM	Papers
Style-breach detection	word	Word frequencies	[179]
Style-breach detection	word	Vectors of lexical and syntactic features	[131, 134]
Style-breach detection	sentence	Vectors of word embeddings	[214]
Style-breach detection	sentence	Vectors of lexical features	[229]
Style-breach detection	sliding window	Vectors of lexical features	[209]
Author clustering	document	Vectors of lexical features	[52, 123, 141, 166, 255]
Author clustering	document	Word frequencies	[139]
Author clustering	document	Word embeddings	[220]
Author verification	document	Word frequencies	[159]
Author verification	document	Vectors of lexical features	[103, 121, 122, 140]
Author verification	document	Vectors of lexical and syntactic features	[44, 108, 132, 193]
Author verification	document	Vectors of syntactic features	[196]

Syntax-based Methods

Syntax-based detection methods typically operate on the sentence level and employ PoS tagging to determine the syntactic structure of sentences [99, 245]. The syntactic information helps to address morphological ambiguity during the lemmatization or stemming step of preprocessing [117], or to reduce the workload of a subsequent semantic analysis, typically by exclusively comparing the pairs of words belonging to the same PoS class [102]. Many intrinsic detection methods use the frequency of PoS tags as a stylometric feature. The method of Tschuggnall and Specht [245] relies solely on the syntactic structure of sentences. Table 11 presents an overview of papers using syntax-based methods.

Table 9. Detection Methods Querying Web Search Engines

Strategy	Paper
Querying the words with the highest tf-idf value	[89, 136, 145, 205, 207, 238]
Querying the least frequent words	[106, 155]
Querying the least frequent strings	[152]
Querying the words with the highest tf-idf value as well as noun phrases	[63, 64, 278]
Querying the nouns and most frequent words	[203]
Querying the nouns and verbs	[143]
Querying the nouns, verbs, and adjectives	[207, 268, 269, 270]
Querying the nouns, facts (dates, names, etc.) as well as the most frequent words	[49]
Querying keywords and the longest sentence in a paragraph	[237, 239]
Comparing different querying heuristics	[133]
Incrementing passage length and passage selection heuristics	[257]
Query expansion by words from UMLS Meta-thesaurus	[176]

Table 10. Detection Methods Employing the General Impostors Method

Task	Papers
Author verification	[103, 135, 146, 171, 224]

Table 11. Syntax-based Detection Methods

Approach	Method	Purpose / Method	Papers
Extrinsic	PoS tagging	Addressing morphological ambiguity	[117, 118]
		Word comparisons within the same PoS class only	[102, 168]
		Combined with stop-words	[101]
		Comparing PoS sequences	[272]
		Combination with PPM compression	[112]
Intrinsic	PoS tags as stylometric features	PoS frequency	[115, 149, 157, 184, 276]
		PoS n-gram frequency	[4, 30, 44, 87, 135, 170, 196, 258]
		PoS frequency, PoS n-gram frequency, starting PoS tag	[158]
	Comparing syntactic trees	Direct comparison	[245, 246]
		Integrated syntactic graphs	[99]

Semantics-based Methods

Papers presenting semantics-based detection methods are the largest group in our collection. This finding reflects the importance of detecting obfuscated forms of academic plagiarism, for which semantics-based detection methods are the most promising approach [216]. Semantics-based methods operate on the hypothesis that the semantic similarity of two passages depends on the occurrence of similar semantic units in these passages. The semantic similarity of two units derives from their occurrence in similar contexts.

Many semantics-based methods use thesauri (e.g., WordNet or EuroVoc[7]). Including semantic features, like synonyms, hypernyms, and hyponyms, in the analysis improves the performance of paraphrase identification [9]. Using a canonical synonym for each word helps detecting synonym-replacement obfuscation and reduces the vector space dimension [206]. Sentence segmentation and text tokenization are crucial parameters for all semantics-based detection methods. Tokenization extracts the atomic units of the analysis, which are typically either words or phrases. Most papers in our collection use words as tokens.

Employing established semantic text analysis methods like Latent Semantic Analysis (LSA), Explicit Semantic Analysis (ESA), and word embeddings for extrinsic plagiarism detection is a popular and successful approach. This group of methods follows the idea of "distributional semantics," i.e., terms co-occurring in similar contexts tend to convey a similar meaning. In the reverse conclusion, distributional semantics assumes that similar distributions of terms indicate semantically similar texts. The methods differ in the scope within which they consider co-occurring terms. Word embeddings consider only the immediately surrounding terms, LSA analyzes the entire document and ESA uses an external corpus.

Latent Semantic Analysis is a technique to reveal and compare the underlying semantic structure of texts [55]. To determine the similarity of term distributions in texts, LSA computes a matrix, in which rows represent terms, columns represent documents and the entries of the matrix typically represent log-weighted tf-idf values [46]. LSA then employs Singular Value Decomposition (SVD) or similar dimensionality reduction techniques to find a lower-rank approximation of the term-document matrix by reducing the number of rows (i.e., pruning less relevant terms) while maintaining the similarity distribution between columns (i.e., the text representations). The terms remaining after the dimensionality reduction are assumed to be most representative of the semantic meaning of the text. Hence, comparing the rank-reduced matrix-representations of texts allows computing the semantic similarity of the texts [46].

LSA can reveal similarities between texts that traditional vector space models cannot express [116]. The ability of LSA to address synonymy is beneficial for paraphrase identification. For example, Satyapanich et al. [222] considered two sentences as paraphrases if their LSA similarity is above a threshold. While LSA performs well in addressing synonymy, its ability to reflect polysemy is limited [55].

Ceska [46] first applied LSA for plagiarism detection. AlSallal et al. [15] proposed a novel weighting approach that assigns higher weights to the most common terms and used LSA as a stylometric feature for intrinsic plagiarism detection. Aldarmaki and Diab [11] used weighted matrix factorization—a method similar to LSA—for cross-language paraphrase identification. Table 12 lists other papers employing LSA for extrinsic and intrinsic plagiarism detection.

Explicit Semantic Analysis is an approach to model the semantics of a text in a high-dimensional vector space of semantic concepts [82]. Semantic concepts are the topics in a man-made knowledge base corpus (typically Wikipedia or other encyclopedias). Each article in the knowledge base is an explicit description of the semantic content of the concept, i.e., the topic of the article [163]. ESA builds a "semantic interpreter" that allows representing texts as concept vectors whose components reflect the relevance of the text for each of the semantic concepts, i.e., knowledge base articles [82]. Applying vector similarity measures, such as the cosine metric, to the concept vectors then allows determining the texts' semantic similarity.

Table 13 shows detection methods that employed ESA depending on the corpus used to build the semantic interpreter. Constructing the semantic interpreter from multilingual corpora, such as Wikipedia, allows the application of ESA for cross-language plagiarism detection [78]. ESA has

[7] https://publications.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc.

Table 12. Detection Methods Employing LSA

Approach	Task	Extension of LSA	Papers
Extrinsic	Document-level detection	LSA with phrase tf-idf	[117, 118]
		LSA in combination with other methods	[256]
	Candidate retrieval	LSA only	[230]
	Paraphrase identification	LSA only	[222]
		LSA with machine learning	[13, 243, 259, 275]
		Weighted matrix factorization	[11]
Intrinsic	Document-level detection	LSA with stylometric features	[14]
	Author identification	LSA with machine learning	[15, 43]
		LSA at CNG level	[221]

Table 13. Detection Methods Employing ESA

Corpus	Papers
Wikipedia (monolingual)	[124, 164, 260]
Wikipedia (cross-language)	[74, 78]
Wikipedia + FanFiction	[174]

Table 14. Detection Methods Employing IR-based Semantic Similarity

Document sets	Papers
Articles from Wikipedia	[120]
Synonyms from Farsnet	[206]

several applications beyond PD, e.g., when applied for document classification, ESA achieved a precision above 95% [124, 174].

The *Information Retrieval-based semantic similarity approach* proposed by Itoh [120] is a generalization of ESA. The method models a text passage as a set of words and employs a Web search engine to obtain a set of relevant documents for each word in the set. The method then computes the semantic similarity of the text passages as the similarity of the document sets obtained, typically using the Jaccard metric. Table 14 presents papers that also follow this approach.

Word embeddings is another semantic analysis approach that is conceptually related to ESA. While ESA considers term occurrences in each document of the corpus, word embeddings exclusively analyze the words that surround the term in question. The idea is that terms appearing in proximity to a given term are more characteristic of the semantic concept represented by the term in question than more distant words. Therefore, terms that frequently co-occur in proximity within texts should also appear closer within the vector space [73]. In cross-language plagiarism detection, word embeddings outperformed other methods when syntactic weighting was employed [73]. Table 15 summarizes papers that employ word embeddings.

Word Alignment is a semantic analysis approach widely used for machine translation [240] and paraphrase identification. Words are aligned, i.e., marked as related, if they are semantically similar. Semantic similarity of two words is typically retrieved from an external database, like WordNet. The semantic similarity of two sentences is then computed as the proportion of aligned words. Word alignment approaches achieved the best performance for the paraphrase identification task at SemEval 2014 [240] and were among the top-performing approaches at SemEval-2015 [9, 242].

Table 15. Detection Methods Employing Word Embeddings

Approach	Task	Papers
Extrinsic	Candidate retrieval	[175]
	Cross-language PD	[95]
Intrinsic	Paraphrase identification	[23, 72, 73, 113, 243]
	Style-breach detection	[214]
	Author clustering	[220]

Table 16. Detection Methods Employing Word Alignment and CL-ASA

Method	Papers
Word alignment only	[240, 241]
Word alignment-based modification of Jaccard and Levenshtein measure	[17]
Word alignment in combination with machine learning	[111, 242, 271]
CL-ASA	[28, 74]
Translation + word alignment	[72]

Cross-language alignment-based similarity analysis (CL-ASA) is a variation of the word alignment approach for cross-language semantic analysis. The approach uses a parallel corpus to compute the similarity that a word x in the suspicious document is a valid translation of the term y in a potential source document for all terms in the suspicious and the source documents. The sum of the translation probabilities yields the probability that the suspicious document is a translation of the source document [28]. Table 16 presents papers using Word alignment and CL-ASA.

Graph-based Semantic Analysis. Knowledge graph analysis (KGA) represents a text as a weighted directed graph, in which the nodes represent the semantic concepts expressed by the words in the text and the edges represent the relations between these concepts [79]. The relations are typically obtained from publicly available corpora, such as BabelNet[8] or WordNet. Determining the edge weights is the major challenge in KGA. Traditionally, edge weights were computed from analyzing the relations between concepts in WordNet [79]. Salvador et al. [79] improved the weighting procedure by using continuous skip-grams that additionally consider the context in which the concepts appear. Applying graph similarity measures yields a semantic similarity score for documents or parts thereof (typically sentences).

Inherent characteristics of KGA like word sense disambiguation, vocabulary expansion, and language independence are highly beneficial for plagiarism detection. Thanks to these characteristics, KGA is resistant to synonym replacements and syntactic changes. Using multilingual corpora allows the application of KGA for cross-language PD [79]. KGA achieves high detection effectiveness if the text is translated literally; for paraphrased translations, the results are worse [77].

The *universal networking language approach* proposed by Avishek and Bhattacharyyan [53] is conceptually similar to KGA. The method constructs a dependency graph for each sentence and then compares the lexical, syntactic, and semantic similarity separately. Kumar [147] used *semantic graphs* for the seeding phase of the detailed analysis stage. In those graphs, the nodes corresponded to all words in a document or passage. The edges represented the adjacency of the words. The edge weights expressed the semantic similarity of words based on the probability that the words occur in a 100-word window within a corpus of DBpedia[9] articles. Overlapping passages in two documents were identified using the minimum weight bipartite clique cover.

[8] https://babelnet.org/.
[9] https://wiki.dbpedia.org/.

Table 17. Detection Methods Employing Graph-based Analysis

Task	Method	Papers
Document-level detection	Knowledge graph analysis	[79]
Detailed analysis	Semantic graphs	[147]
Detailed analysis	Word n-gram graphs for sentences	[169]
Paraphrase identification	Knowledge graph analysis	[168]
Paraphrase identification	Universal networking language	[53]
Cross-language plagiarism detection	Knowledge graph analysis	[76, 77, 78]

Table 18. Detection Approaches Employing Semantic Role Labeling

Task	Papers
Document-level detection	[182, 188]
Paraphrase identification	[71]

Table 19. Overview of Idea-based Detection Methods

Task	Approach	Papers
Monolingual plagiarism detection	Citation-based PD	[93, 104, 161, 191, 192]
	Math-based PD	[164, 165]
	Image-based PD	[162]
Cross-lingual plagiarism detection	CbPD	[94]

Table 17 presents detection methods that employ graph-based semantic analysis.

Semantic Role Labeling (SRL) determines the semantic roles of terms in a sentence, e.g., the subject, object, events, and relations between these entities, based on roles defined in linguistic resources, such as PropBank[10] or VerbNet.[11] The goal is to extract "who" did "what" to "whom" "where" and "when" [188]. The first step in SRL is PoS tagging and syntax analysis to obtain the dependency tree of a sentence. Subsequently, the semantic annotation is performed [71].

Paul and Jamal [188] used SRL in combination with sentence ranking for document-level plagiarism detection. Hamza and Salim [182] employed SRL to extract arguments from sentences, which they used to quantify and compare the syntactic and semantic similarity of the sentences. Ferreira et al. [71] obtained the similarity of sentences by combining various features and measures using machine learning. Table 18 lists detection approaches that employ SRL.

Idea-based Methods

Idea-based methods analyze non-textual content elements to identify obfuscated forms of academic plagiarism. The goal is to complement detection methods that analyze the lexical, syntactic, and semantic similarity of text to identify plagiarism instances that are hard to detect both for humans and for machines. Table 19 lists papers that proposed idea-based detection methods.

Citation-based plagiarism detection (CbPD) proposed by Gipp et al. [91] analyses patterns of in-text citations in academic documents, i.e., identical citations occurring in proximity or in a similar order within two documents. The idea is that in-text citations encode semantic information language-independently. Thus, analyzing in-text citation patterns can indicate shared structural

[10] http://verbs.colorado.edu/~mpalmer/projects/ace.html.
[11] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

and semantic similarity among texts. Assessing semantic and structural similarity using citation patterns requires significantly less computational effort than approaches for semantic and syntactic text analysis [90]. Therefore, CbPD is applicable for the candidate retrieval and the detailed analysis stage [161] of monolingual [90, 93] and cross-lingual [92] detection methods. For weakly obfuscated instances of plagiarism, CbPD achieved comparable results as lexical detection methods; for paraphrased and idea plagiarism, CbPD outperformed lexical detection methods in the experiments of Gipp et al. [90, 93]. Moreover, the visualization of citation patterns was found to facilitate the inspection of the detection results by humans, especially for cases of structural and idea plagiarism [90, 93]. Pertile et al. [191] confirmed the positive effect of combining citation and text analysis on the detection effectiveness and devised a hybrid approach using machine learning. CbPD can also alert a user when the in-text citations are inconsistent with the list of references. Such inconsistency may be caused by mistake, or deliberately to obfuscate plagiarism.

Meuschke et al. [163] proposed *mathematics-based plagiarism detection* (MathPD) as an extension of CbPD for documents in the Science, Technology, Engineering and Mathematics (STEM) fields. Mathematical expressions share many properties of academic citations, e.g., they are essential components of academic STEM documents, are language-independent, and contain rich semantic information. Furthermore, some disciplines, such as mathematics and physics, use academic citations sparsely [167]. Therefore, a citation-based analysis alone is less likely to reveal suspicious content similarity for these disciplines [163], [165]. Meuschke et al. showed that an exclusive math-based similarity analysis performed well for detecting confirmed cases of academic plagiarism in STEM documents [163]. Combining a math-based and a citation-based analysis further improved the detection performance for confirmed cases of plagiarism [165].

Image-based plagiarism detection analyze graphical content elements. While a large variety of methods to retrieve similar images have been proposed [56], few studies investigated the application of content-based image retrieval approaches for academic plagiarism detection. Meuschke et al. [162] is the only such study we encountered during our data collection. The authors proposed a detection approach that integrates established image retrieval methods with novel similarity assessments for images that are tailored to plagiarism detection. The approach has been shown to retrieve both copied and altered figures.

Ensembles of Detection Methods

Each class of detection methods has characteristic strengths and weaknesses. Many authors showed that combining detection methods achieves better results than applying the methods individually [7, 62, 78, 128, 133, 234, 242, 273, 275]. By assembling the best-performing detection methods in PAN 2014, the organizers of the workshop created a meta-system that performed best overall [232].

In intrinsic plagiarism detection, combining feature analysis methods is a standard approach [233], since an author's writing style always comprises of a multitude of stylometric features [127]. Many recent author verification methods employ machine learning to select the best performing feature combination [234].

In general, there are three ways of combining plagiarism detection methods:

- Using *adaptive algorithms* that determine the obfuscation strategy, choose the detection method, and set similarity thresholds accordingly
- Using an *ensemble of detection methods* whose results are combined using static weights
- Using *machine learning* to determine the best-performing combination of detection methods

The winning approach at PAN 2014 and 2015 [216] used an *adaptive algorithm*. After finding the seeds of overlapping passages, the authors extended the seeds using two different thresholds for

Table 20. Ensembles of Detection Methods

Task	Method	Type of ensemble	Papers
Document-level detection		Linguistic knowledge	[2]
Candidate retrieval	Querying a Web search engine	Combination of querying heuristics	[133]
Detailed analysis	Vector space model	Adaptive algorithm	[186, 216, 217]

the maximum gap. Based on the length of the passages, the algorithm automatically recognized different plagiarism forms and set the parameters for the VSM-based detection method accordingly.

The "*linguistic knowledge approach*" proposed by Abdi et al. [2] exemplifies an *ensemble of detection methods*. The method combines the analysis of syntactic and semantic sentence similarity using a linear combination of two similarity metrics: (i) the cosine similarity of semantic vectors and (ii) the similarity of syntactic word order vectors [2]. The authors showed that the method outperformed other contesters on the PAN-10 and PAN-11 corpora. Table 20 lists other ensembles of detection methods.

Machine Learning approaches for plagiarism detection typically train a classification model that combines a given set of features. The trained model can then be used to classify other datasets. Support vector machine (SVM) is the most popular model type for plagiarism detection tasks. SVM uses statistical learning to minimize the distance between a hyperplane and the training data. Choosing the hyperplane is the main challenge for correct data classification [66].

Machine-learning approaches are very successful in intrinsic plagiarism detection. Supervised machine-learning methods, specifically random forests, were the best-performing approach at the intrinsic detection task of the PAN 2015 competition [233]. The best-known method for author verification is *unmasking* [232], which uses an SVM classifier to distinguish the stylistic features of the suspicious document from a set of documents for which the author is known. The idea of unmasking is to train and run the classifier and then remove the most significant features of the classification model and rerun the classification. If the classification accuracy drops significantly, then the suspicious and known documents are likely from the same author; otherwise, they are likely written by different authors [232]. There is no consensus on the stylometric features that are most suitable for authorship identification [158]. Table 21 gives an overview of intrinsic detection methods that employ machine-learning techniques.

For extrinsic plagiarism detection, the application of machine learning has been studied for various components of the detection process [208]. Gharaviet al. [88] used machine learning to determine the suspiciousness thresholds for a vector space model. Zarrella et al. [273] won the SemEval competition in 2015 with their ensemble of seven algorithms; most of them used machine learning. While Hussain and Suryani [116] successfully used an SVM classifier for the candidate retrieval stage [269], Williams et al. compared many supervised machine-learning methods and concluded that applying them for classifying and ranking Web search engine results did not improve candidate retrieval. Kanjirangat and Gupta [252] used a genetic algorithm to detect idea plagiarism. The method randomly chooses a set of sentences as chromosomes. The sentence sets that are most descriptive of the entire document are combined and form the next generation. In this way, the method gradually extracts the sentences that represent the idea of the document and can be used to retrieve similar documents.

Sánchez-Vega et al. [218] proposed a method termed *rewriting index* that evaluates the degree of membership of each sentence in the suspicious document to a possible source document. The method uses five different Turing machines to uncover verbatim copying as well as basic transformations on the word level (insertion, deletion, substitution). The output values of the Turing machines are used as the features to train a Naïve Bayes classifier and identify reused passages.

Table 21. Machine-learning-based Intrinsic Plagiarism Detection Methods

Task	Classifier	Features	Papers
Style-breach detection	Gradient Boosting Regression Trees	Lexical, syntax	[149]
Author identification	SVM	Semantic (LSA)	[15]
Author clustering	Recurrent ANN	Lexical	[26],
	SVM	Lexical, syntax	[276]
Author verification	Recurrent ANN	Lexical	[25]
	k-nearest neighbor	Lexical	[109]
		Lexical, syntax	[87]
	Homotopy-based classification	Lexical	[103]
	Naïve Bayes	Lexical	[159]
	SVM	Lexical, syntax	[4, 70, 115, 146, 258]
	Equal error rate	Lexical	[110]
	Decision Tree	Lexical	[81]
	Random Forest	Lexical, syntax	[30, 158, 184]
	Genetic algorithm	Lexical, syntax	[170, 171]
	Multilayer perceptron	Lexical, semantic (LSA)	[16]
	Many	Lexical	[151, 219]
		Lexical, syntax	[172]

In the approach of Afzal et al. [5], the linear combination of supervised and unsupervised machine-learning methods outperformed each of the methods applied individually. In the experiments of Alfikri and Purwarianti [13], SVM classifiers outperformed Naïve Bayes classifiers. In the experiments of Subroto and Selamat [236], the best performing configuration was a hybrid model that combined SVM and an artificial neural network (ANN). El-Alfy et al. [62] found that an abductive network outperformed SVM. However, as shown in Table 22, SVM is the most popular classifier for extrinsic plagiarism detection methods. Machine learning appears to be more beneficial when applied for the detailed analysis, as indicated by the fact that most extrinsic detection methods apply machine learning for that stage (cf. Table 22).

Evaluation of Plagiarism Detection Methods

The availability of datasets for development and evaluation is essential for research on natural language processing and information retrieval. The PAN series of benchmark competitions is a comprehensive and well–established platform for the comparative evaluation of plagiarism detection methods and systems [197]. The PAN test datasets contain artificially created monolingual (English, Arabic, Persian) and—to a lesser extent—cross-language plagiarism instances (German and Spanish to English) with different levels of obfuscation. The papers included in this review that present lexical, syntactic, and semantic detection methods mostly use PAN datasets[12] or the Microsoft Research Paraphrase corpus.[13] Authors presenting idea-based detection methods that analyze non-textual content features or cross-language detection methods for non-European languages typically use self-created test collections, since the PAN datasets are not suitable for these tasks. A comprehensive review of corpus development initiatives is out of the scope of this article.

[12] https://pan.webis.de/data.html.
[13] https://www.microsoft.com/en-us/download/details.aspx?id=52398.

Table 22. Machine-learning-based Extrinsic Plagiarism Detection Methods

Task	Classifier	Features	Papers
Document-level detection	SVM	Semantic	[66, 116]
	SVM, Naïve Bayes	Lexical, semantic	[13]
	Decision tree, k-nearest neighbor	Syntax	[38]
	Naïve Bayes, SVM, Decision tree	Lexical, syntax	[254]
	Many	Semantic (CbPD)	[191]
Candidate retrieval	SVM	Lexical	[142]
	Linear discriminant analysis	Lexical, syntax	[270]
	Genetic algorithm	Lexical, syntax	[252]
Detailed analysis	Logical regression model	Lexical, syntax, semantic	[144]
	Naïve Bayes	Lexical	[218]
	Naïve Bayes, Decision Tree, Random Forest	Lexical	[129]
	SVM	Lexical, semantic	[147]
Paraphrase identification	SVM	Lexical	[67]
		Lexical, semantic	[111, 259]
		Lexical, syntax, semantic	[41, 42, 130]
		MT metrics	[36]
		ML with syntax and semantic features	[213]
	k-nearest neighbor, SVM, artificial neural network	Lexical	[236]
	SVM, Random forest, Gradient boosting	Lexical, syntax, semantic, MT metrics	[243]
	SVM, MaxEnt	Lexical, syntax, semantic	[9]
	Abductive networks	Lexical	[62]
	Linear regression	Lexical, syntax, semantic	[53]
	L2-regularized logistic regression	Lexical, syntax, semantic, ML	[273]
	Ridge regression	Lexical, semantic	[189]
	Gaussian process regression	Lexical, semantic	[194]
	Isotonic regression	Semantic	[153]
	Artificial neural network	Lexical, semantic	[113]
	Deep neural network	Syntax, semantic	[5]
		Semantic	[6]
	Decision Tree	Semantic	[23]
		Lexical, syntax, semantic	[72, 73]
	Random Forest	Semantic, MT metrics	[204]
	Many	Lexical, semantic	[71, 215]
		Lexical, syntax, semantic	[262, 275]
Cross-language PD	Artificial neural networks	Semantic	[78]

Since plagiarism detection is an information retrieval task, precision, recall, and F–measure are typically employed to evaluate plagiarism detection methods. A notable use-case-specific extension of these general performance measures is the PlagDet metric. Potthast et al. introduced the metric to evaluate the performance of methods for the detailed analysis stage in external plagiarism detection [201]. A method may detect only a fragment of a plagiarism instance or report a coherent instance as multiple detections. To account for these possibilities, Potthast et al. included the granularity score as part of the PlagDet metric. The granularity score is the ratio of the detections a method reports and the true number of plagiarism instances.

PLAGIARISM DETECTION SYSTEMS

Plagiarism detection systems implement (some of) the methods described in the previous sections. To be applicable in practice, the systems must address the tradeoff between detection performance and processing speed [102], i.e., find sources of plagiarism with reasonable computational costs.

Most systems are Web-based; some can run locally. The systems typically highlight the parts of a suspicious document that likely originate from another source as well as which source that is. Understanding *how* the source was changed is often left to the user. Providers of plagiarism detection systems, especially of commercial systems, rarely publish information on the detection methods they employ [85, 256]. Thus, estimating to what extent plagiarism detection research influences practical applications is difficult.

Velásquez et al. [256] provided a text-matching software and described its functionality that included the recognition of quotes. The system achieved excellent results in the PAN 10 and PAN 11 competitions. Meanwhile, the authors commercialized the system [195].

Academics and practitioners are naturally interested in which detection system achieves the best results. Weber-Wulff and her team performed the most methodologically sound investigation of this question in 2004, 2007, 2008, 2010, 2011, 2012, and 2013 [266]. In their latest benchmark evaluation, the group compared 15 systems using documents written in English and German.

Chowdhury and Bhattacharyya [48] provided an exhaustive list of currently available plagiarism detection systems. Unfortunately, the description of each system is short, and the authors did not provide performance comparisons. Pertile et al. [191] summarized the basic characteristics of 17 plagiarism detection systems. Kanjirangat and Gupta [251] compared four publicly available systems. They used four test documents that contained five forms of plagiarism (copy-and-paste, random obfuscation, translation to Hindi and back, summarization). All systems failed to identify plagiarism instances other than copy-and-paste and random obfuscation.

There is consensus in the literature that the inability of plagiarism detection systems to identify obfuscated plagiarism is currently their most severe limitation [88, 251, 266].

In summary, there is a lack of systematic and methodologically sound performance evaluations of plagiarism detection systems, since the benchmark comparisons of Weber-Wulff ended in 2013. This lack is problematic, since plagiarism detection systems are typically a key building block of plagiarism policies. Plagiarism detection methods and plagiarism policies are the subjects of extensive research. We argue that plagiarism detection systems should be researched just as extensively but are currently not.

DISCUSSION

In this section, we summarize the advancements in the research on methods to detect academic plagiarism that our review identified. Figure 2 depicts the suitability of the methods discussed in the previous sections for identifying the plagiarism forms presented in our typology. As shown in the Figure, n-gram comparisons are well-suited for detecting character-preserving plagiarism and partially suitable for identifying ghostwriting and syntax-preserving plagiarism. Stylometry is

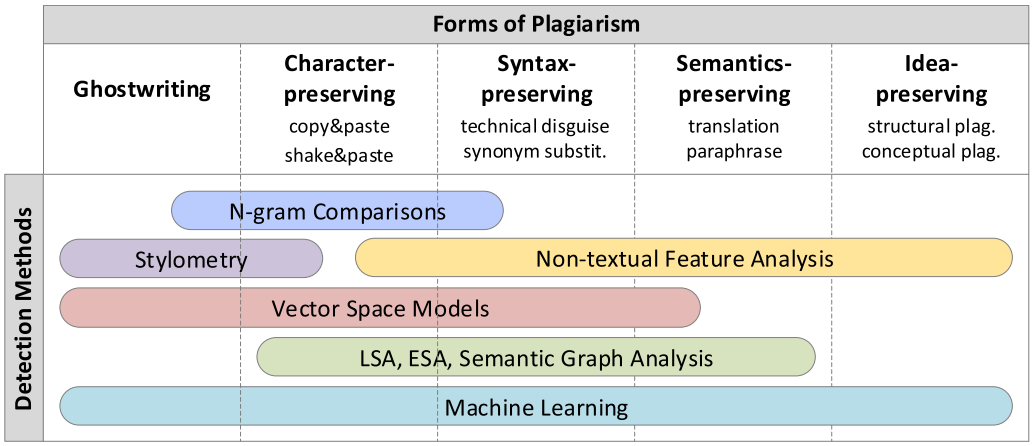

Fig. 2. Suitability of detection methods for identifying certain forms of plagiarism.

routinely applied for intrinsic plagiarism detection and can reveal ghostwriting and copy-and-paste plagiarism. Vector space models have a wide range of applications but appear not to be particularly beneficial for detecting idea plagiarism. Semantics-based methods are tailored to the detection of semantics-preserving plagiarism, yet also perform well for character-preserving and syntax-preserving forms of plagiarism. Non-textual feature analysis and machine learning are particularly beneficial for detecting strongly obfuscated forms of plagiarism, such as semantics-preserving and idea-preserving plagiarism. However, machine learning is a universal approach that also performs well for less strongly disguised forms of plagiarism.

Extrinsic Plagiarism Detection

The first observation of our literature survey is that ensembles of detection methods tend to outperform approaches based on a single method [93, 161]. Chong experimented with numerous methods for preprocessing as well as with shallow and deep NLP techniques [47]. He tested the approaches on both small and large-scale corpora and concluded that a combination of string-matching and deep NLP techniques achieves better results than applying the techniques individually.

Machine-learning approaches represent the logical evolution of the idea to combine heterogeneous detection methods. Since our previous review in 2013, unsupervised and supervised machine-learning methods have found increasingly wide-spread adoption in plagiarism detection research and significantly increased the performance of detection methods. Baroni et al. [27] provided a systematic comparison of vector-based similarity assessments. The authors were particularly interested in whether unsupervised count-based approaches like LSA achieve better results than supervised prediction-based approaches like Softmax. They concluded that the prediction-based methods outperformed their count-based counterparts in precision and recall while requiring similar computational effort. We expect that the research on applying machine learning for plagiarism detection will continue to grow significantly in the future.

Considering the heterogeneous forms of plagiarism (see the typology section), the static one-fits-all approach observable in most plagiarism detection methods before 2013 is increasingly replaced by adaptive detection algorithms. Many recent detection methods first seek to identify the likely obfuscation method and then apply the appropriate detection algorithm [79, 198], or at least to dynamically adjust the parameters of the detection method [216].

Graph-based methods operating on the syntactic and semantic levels achieve comparable results to other semantics-based methods. Mohebbi and Talebpour [168] successfully employed graph-based methods to identify paraphrases. Franco-Salvador et al. [79] demonstrated the suitability of knowledge graph analysis for cross-language plagiarism detection.

Several researchers showed the benefit of analyzing non-textual content elements to improve the detection of strongly obfuscated forms of plagiarism. Gipp et al. demonstrated that analyzing in-text citation patterns achieves higher detection rates than lexical approaches for strongly obfuscated forms of academic plagiarism [90, 92–94]. The approach is computationally modest and reduces the effort required of users for investigating the detection results. Pertile et al. [191] combined lexical and citation-based approaches to improve detection performance. Eisa et al. [61] strongly advocated for additional research on analyzing non-textual content features. The research by Meuschke et al. on analyzing images [162] and mathematical expressions [164] confirms that non-textual detection methods significantly enhance the detection capabilities. Following the trend of combining detection methods, we see the analysis of non-textual content features as a promising component of future integrated detection approaches.

Surprisingly many papers in our collection addressed plagiarism detection for Arabic and Persian texts (e.g., References [22, 118, 231, 262]). The interest in plagiarism detection for the Arabic language led the organizers of the PAN competitions to develop an Arabic corpus for intrinsic plagiarism detection [34]. In 2015, the PAN organizers also introduced a shared task on plagiarism detection for Arabic texts [32], followed by a shared task for Persian texts one year later [22]. While these are promising steps toward improving plagiarism detection for Arabic, Wali et al. [262] noted that the availability of corpora and lexicons for Arabic is still insufficient when compared to other languages. This lack of resources and the complex linguistic features of the Arabic language cause plagiarism detection for Arabic to remain a significant research challenge [262].

For cross-language plagiarism detection methods, Ferrero et al. [74] introduced a five-class typology that still reflects the state of the art: cross-language character n-grams (CL-CNG), cross-language conceptual thesaurus-based similarity (CL-CTS), cross-language alignment-based similarity analysis (CL-ASA), cross-language explicit semantic analysis (CL-ESA), and translation with monolingual analysis (T+MA). Franco-Salvador et al. [80] showed that the performance of these methods varies depending on the language and corpus. The observation that the combination of detection methods improves the detection performance also holds for the cross-language scenario [80]. In the analysis of Ferrero et al. [74], the detection performance of methods exclusively depended on the size of the chosen chunk but not on the language, nor the dataset. Translation with monolingual analysis is a widely used approach. For the cross-language detection task (Spanish–English) at the SemEval competition in 2016, most of the contesters applied a machine translation from Spanish to English and then compared the sentences in English [7]. However, some authors do not consider this approach as cross-language plagiarism detection but as monolingual plagiarism detection with translation as a preprocessing step [80].

Intrinsic Plagiarism Detection

For intrinsic plagiarism detection, authors predominantly use lexical and syntax-based text analysis methods. Widely analyzed lexical features include character n-grams, word frequencies, as well as the average lengths of words, sentences, and paragraphs [247]. The most common syntax-based features include PoS tag frequencies, PoS tag pair frequencies, and PoS structures [247]. At the PAN competitions, methods that analyzed lexical features and employed simple clustering algorithms achieved the best results [200].

For the author verification task, the most successful methods treated the problem as a binary classification task. They adopted the extrinsic verification paradigm by using texts from other

authors to identify features that are characteristic of the writing style of the suspected author [233]. The general impostors method is a widely used and largely successful realization of this approach [135, 146, 159, 224].

From a practitioner's perspective, intrinsic detection methods exhibit several shortcomings. First, stylometric comparisons are inherently error-prone for documents collaboratively written by multiple authors [209]. This shortcoming is particularly critical, since most scientific publications have multiple authors [39]. Second, intrinsic methods are not well suited for detecting paraphrased plagiarism, i.e., instances in which authors illegitimately reused content from other sources that they presented in their own words. Third, the methods are generally not reliable enough for practical applications yet. Author identification methods achieve a precision of approximately 60%, author profiling methods of approximately 80% [200]. These values are sufficient for raising suspicion and encouraging further examination but not for proving plagiarism or ghostwriting. The availability of methods for automated author obfuscation aggravates the problem. The most effective methods can mislead the identification systems in almost half of the cases [199]. Fourth, intrinsic plagiarism detection approaches cannot point an examiner to the source document of potential plagiarism. If a stylistic analysis raised suspicion, then extrinsic detection methods or other search and retrieval approaches are necessary to discover the potential source document(s).

Other Applications of Plagiarism Detection Methods

Aside from extrinsic and intrinsic plagiarism detection, the methods described in this article have numerous other applications such as machine translation [67], author profiling for marketing applications [211], spam detection [248], law enforcement [127, 211], identifying duplicate accounts in internet fora [4], identifying journalistic text reuse [47], patent analysis [1], event recognition based on tweet similarity [24, 130], short answer scoring based on paraphrase identification [242], or native language identification [119].

CONCLUSION

In 2010, Mozgovoy et al. [173] proposed a roadmap for the future development of plagiarism detection systems. They suggested the inclusion of syntactic parsing, considering synonym thesauri, employing LSA to discover "tough plagiarism," intrinsic plagiarism detection, and tracking citations and references. As our review of the literature shows, all these suggestions have been realized. Moreover, the field of plagiarism detection has made a significant leap in detection performance thanks to machine learning.

In 2015, Eisa et al. [61] praised the effort invested into improving text-based plagiarism detection but noted a critical lack of *"techniques capable of identifying plagiarized figures, tables, equations and scanned documents or images."* While Meuschke et al. [163, 165] proposed initial approaches that addressed these suggestions and achieved promising results, most of the research still addresses text-based plagiarism detection only.

A generally observable trend is that approaches that integrate different detection methods—often with the help of machine learning—achieve better results. In line with this observation, we see a large potential for the future improvement of plagiarism detection methods in integrating non-textual analysis approaches with the many well-performing approaches for the analysis of lexical, syntactic, and semantic text similarity.

To summarize the contributions of this article, we refer to the four questions Kitchenham et al. [138] suggested to assess the quality of literature reviews:

1. *"Are the review's inclusion and exclusion criteria described and appropriate?*
2. *Is the literature search likely to have covered all relevant studies?*

3. *Did the reviewers assess the quality/validity of the included studies?*
4. *Were the basic data/studies adequately described?"*

We believe that the answers to these four questions are positive for our survey. Our article summarizes previous research and identifies research gaps to be addressed in the future. We are confident that this review will help researchers newly entering the field of academic plagiarism detection to get oriented as well that it will help experienced researchers to identify related works. We hope that our findings will aid in the development of more effective and efficient plagiarism detection methods and system that will then facilitate the implementation of plagiarism policies.

REFERENCES

[1] Assad Abbas, Limin Zhang, and Samee U. Khan. 2014. A literature review on the state-of-the-art in patent analysis. *World Pat. Inf.* 37 (2014), 3–13. DOI: 10.1016/j.wpi.2013.12.006

[2] Asad Abdi, Norisma Idris, Rasim M. Alguliyev, and Ramiz M. Aliguliyev. 2015. PDLK: Plagiarism detection using linguistic knowledge. *Expert Syst. Appl.* 42, 22 (2015), 8936–8946. DOI: 10.1016/j.eswa.2015.07.048

[3] Samira Abnar, Mostafa Dehghani, Hamed Zamani, and Azadeh Shakery. 2014. Expanded n-grams for semantic text alignment—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[4] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger finder: Taking stylometry to the underground. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy*. 212–226.

[5] Naveed Afzal, Yanshan Wang, and Hongfang Liu. 2016. MayoNLP at SemEval-2016 Task 1: Semantic textual similarity based on lexical semantic net and deep learning semantic model. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 674–679.

[6] Basant Agarwal, Heri Ramampiaro, Helge Langseth, and Massimiliano Ruocco. 2018. A deep network model for paraphrase detection in short text messages. *Inf. Process. Manag.* 54, 6 (2018), 922–937. DOI: 10.1016/j.ipm.2018.06.005

[7] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 497–511.

[8] Mayank Agrawal and Dilip Kumar Sharma. 2016. A state of art on source code plagiarism detection. In *Proceedings of the 2016 2nd International Conference on Next Generation Computing Technologies (NGCT'16)*. 236–241. DOI: 10.1109/NGCT.2016.7877421

[9] Mohammad Al-Smadi, Zain Jaradat, Mahmoud Al-Ayyoub, and Yaser Jararweh. 2017. Paraphrase identification and semantic text similarity analysis in arabic news tweets using lexical, syntactic, and semantic features. *Inf. Process. Manag.* 53, 3 (2017), 640–652. DOI: 10.1016/j.ipm.2017.01.002

[10] Houda Alberts. 2017. Author clustering with the aid of a simple distance measure—Notebook for PAN at CLEF 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.

[11] Hanan Aldarmaki and Mona Diab. 2016. GWU NLP at SemEval-2016 Shared Task 1: Matrix factorization for crosslingual STS. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 663–667.

[12] Mahmoud Alewiwi, Cengiz Orencik, and Erkay Savas. 2016. Efficient top-k similarity document search utilizing distributed file systems and cosine similarity. *Cluster Comput.* 19, 1 (2016), 109–126. DOI: 10.1007/s10586-015-0506-0

[13] Zakiy Firdaus Alfikri and Ayu Purwarianti. 2014. Detailed analysis of extrinsic plagiarism detection system using machine learning approach (naive bayes and svm). *Indones. J. Electr. Eng. Comput. Sci.* 12, 11 (2014), 7884–7894.

[14] Muna Alsallal, Rahat Iqbal, Saad Amin, and Anne James. 2013. Intrinsic plagiarism detection using latent semantic indexing and stylometry. In *Proceedings of the 2013 6th International Conference on Developments in eSystems Engineering*. 145–150. DOI: 10.1109/DeSE.2013.34

[15] Muna AlSallal, Rahat Iqbal, Saad Amin, Anne James, and Vasile Palade. 2016. An integrated machine learning approach for extrinsic plagiarism detection. In *Proceedings of the 2016 9th International Conference on Developments in eSystems Engineering (DeSE'16)*. 203–208. DOI: 10.1109/DeSE.2016.1

[16] Muna AlSallal, Rahat Iqbal, Vasile Palade, Saad Amin, and Victor Chang. 2019. An integrated approach for intrinsic plagiarism detection. *Fut. Gener. Comput. Syst.* 96 (2019), 700–712. DOI: 10.1016/j.future.2017.11.023

[17] Miguel A. Álvarez-Carmona, Marc Franco-Salvador, Esaú Villatoro-Tello, Manuel Montes-y-Gómez, Paolo Rosso, and Luis Villaseñor-Pineda. 2018. Semantically-informed distance and similarity measures for paraphrase plagiarism identification. *J. Intell. Fuzzy Syst.* 34, 5 (2018), 2983–2990.

[18] Faisal Alvi, Mark Stevenson, and Paul Clough. 2014. Hashing and merging heuristics for text reuse detection. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*. 939–946.
[19] Faisal Alvi, Mark Stevenson, and Paul Clough. 2017. Plagiarism detection in texts obfuscated with homoglyphs. In *Advances in Information Retrieval*. 669–675.
[20] Salha Alzahrani. 2015. Arabic plagiarism detection using word correlation in N-Grams with K-Overlapping approach—Working notes for PAN-AraPlagDet at FIRE 2015. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'15)*.
[21] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Trans. Syst. Man, Cybern. C Appl. Rev.* 42, 2 (2012), 133–149.
[22] Habibollah Asghari, Salar Mohtaj, Omid Fatemi, Heshaam Faili, Paolo Rosso, and Martin Potthast. 2016. Algorithms and corpora for persian plagiarism detection. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'16)*. 61.
[23] Duygu Ataman, Jose G. C. De Souza, Marco Turchi, and Matteo Negri. 2016. FBK HLT-MT at SemEval-2016 Task 1: Cross-lingual semantic similarity measurement using quality estimation features and compositional bilingual word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 570–576.
[24] Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Comput. Intell.* 31, 1 (2015), 132–164. DOI: 10.1111/coin.12017
[25] Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.
[26] Douglas Bagnall. 2016. Authorship clustering using multi-headed recurrent neural networks—Notebook for PAN at CLEF 2016. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'16)*.
[27] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 238–247.
[28] Alberto Barrón-Cedeño, Parth Gupta, and Paolo Rosso. 2013. Methods for cross-language plagiarism detection. *Knowl.-Based Syst.* 50 (2013), 211–217. DOI: 10.1016/j.knosys.2013.06.018
[29] Alberto Barrón-Cedeño, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013. Plagiarism meets paraphrasing: insights for the next generation in automatic plagiarism detection. *Comput. Linguist.* 39, 4 (2013), 917–947. DOI: 10.1162/COLI_a_00153
[30] Alberto Bartoli, Alex Dagri, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. 2015. An author verification approach based on differential features—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.
[31] Jeffrey Beall. 2016. Best practices for scholarly authors in the age of predatory journals. *Ann. R. Coll. Surg. Engl.* 98, 2 (2016), 77–79.
[32] Imene Bensalem, Imene Boukhalfa, Paolo Rosso, Lahsen Abouenour, Kareem Darwish, and Salim Chikhi. 2015. Overview of the AraPlagDet PAN@FIRE2015 shared task on arabic plagiarism detection. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'15)*.
[33] Imene Bensalem, Salim Chikhi, and Paolo Rosso. 2013. Building arabic corpora from wikisource. In *Proceedings of the 2013 ACS International Conference on Computer Systems and Applications (AICCSA'13)*. 1–2. DOI: 10.1109/AICCSA.2013.6616474
[34] Imene Bensalem, Paolo Rosso, and Salim Chikhi. 2013. A new corpus for the evaluation of arabic intrinsic plagiarism detection. In *Information Access Evaluation: Multilinguality, Multimodality, and Visualization*. 53–58.
[35] Imene Bensalem, Paolo Rosso, and Salim Chikhi. 2014. Intrinsic plagiarism detection using n-gram classes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1459–1464.
[36] Ergun Bicici. 2016. RTM at SemEval-2016 Task 1: Predicting semantic similarity with referential translation machines and related statistics. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 758–764.
[37] Victoria Bobicev. 2013. Authorship detection with PPM—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.
[38] Hadj Ahmed Bouarara, Amine Rahmani, Reda Mohamed Hamou, and Abdelmalek Amine. 2014. Machine learning tool and meta-heuristic based on genetic algorithms for plagiarism detection over mail service. In *Proceedings of the 2014 IEEE/ACIS 13th International Conference on Computer and Information Science (ICIS'14)*. 157–162. DOI: 10.1109/ICIS.2014.6912125
[39] Barry Bozeman, Daniel Fay, and Catherine P. Slade. 2013. Research collaboration in universities and academic entrepreneurship: The-state-of-the-art. *J. Technol. Transf.* 38, 1 (2013), 1–67. DOI: 10.1007/s10961-012-9281-8
[40] Pearl Brereton, Barbara A. Kitchenham, David Budgen, Mark Turner, and Mohamed Khalil. 2007. Lessons from applying the systematic literature review process within the software engineering domain. *J. Syst. Softw.* 80, 4 (2007), 571–583. DOI: 10.1016/j.jss.2006.07.009

[41] Tomáš Brychcín and Lukáš Svoboda. 2016. UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 588–594.

[42] Davide Buscaldi, Joseph Le Roux, Jorge J. García Flores, and Adrian Popescu. 2013. LIPN-CORE: Semantic text similarity using n-grams, wordnet, syntactic analysis, ESA and information retrieval based features. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*. 63.

[43] Esteban Castillo, Ofelia Cervantes, Darnes Vilariño, David Pinto, and Saul León. 2014. Unsupervised method for the authorship identification task—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[44] Daniel Castro, Yaritza Adame, María Pelaez, and Rafael Muñoz. 2015. Authorship verification, combining linguistic features and different similarity functions—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[45] Daniele Cerra, Mihai Datcu, and Peter Reinartz. 2014. Authorship analysis based on data compression. *Pattern Recogn. Lett.* 42 (2014), 79–84. DOI: 10.1016/j.patrec.2014.01.019

[46] Zdenek Ceska. 2008. Plagiarism detection based on singular value decomposition. In *Advances in Natural Language Processing*. Springer, 108–119.

[47] Man Yan Miranda Chong. 2013. A study on plagiarism detection and plagiarism direction identification using natural language processing techniques. Ph.D. Thesis. University of Wolverhampton.

[48] Hussain A. Chowdhury and Dhruba K. Bhattacharyya. 2016. Plagiarism: Taxonomy, tools and detection techniques. In *Proceedings of the 19th National Convention on Knowledge, Library and Information Networking (NACLIN'16)*.

[49] Daniela Chudá, Jozef Lačný, Maroš Maršalek, Pavel Michalko, and Ján Súkeník. 2013. Plagiarism detection in slovak texts on the web. In *Proceedings of the Conference on Plagiarism across Europe and Beyond*. 249–260.

[50] Guy J. Curtis and Joseph Clare. 2017. How prevalent is contract cheating and to what extent are students repeat offenders? *J. Acad. Ethics* 15, 2 (2017), 115–124. DOI: 10.1007/s10805-017-9278-x

[51] Guy J. Curtis and Lucia Vardanega. 2016. Is plagiarism changing over time? A 10-year time-lag study with three points of measurement. *High. Educ. Res. Dev.* 35, 6 (2016), 1167–1179. DOI: 10.1080/07294360.2016.1161602

[52] Michiel van Dam. 2013. A basic character n-gram approach to authorship verification—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[53] Avishek Dan and Pushpak Bhattacharyya. 2013. Cfilt-core: Semantic textual similarity using universal networking language. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM'13)*. 216–220.

[54] Ali Daud, Wahab Khan, and Dunren Che. 2017. Urdu language processing: a survey. *Artif. Intell. Rev.* 47, 3 (2017), 279–311. DOI: 10.1007/s10462-016-9482-x

[55] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, 6 (1990), 391. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[56] T. Dharani and I. Laurence Aroquiaraj. 2013. A survey on content based image retrieval. In *Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*. 485–490. DOI: 10.1109/ICPRIME.2013.6496719

[57] Michal Ďuračík, Emil Kršák, and Patrik Hrkút. 2017. Current trends in source code analysis, plagiarism detection and issues of analysis big datasets. *Proc. Eng.* 192 (2017), 136–141. DOI: 10.1016/j.proeng.2017.06.024

[58] Nava Ehsan and Azadeh Shakery. 2016. Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information. *Inf. Process. Manag.* 52, 6 (2016), 1004–1017. DOI: 10.1016/j.ipm.2016.04.006

[59] Nava Ehsan and Azadeh Shakery. 2016. A pairwise document analysis approach for monolingual plagiarism detection. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'16)*. 145–148.

[60] Nava Ehsan, Frank Wm. Tompa, and Azadeh Shakery. 2016. Using a dictionary and n-gram alignment to improve fine-grained cross-language plagiarism detection. In *Proceedings of the 2016 ACM Symposium on Document Engineering (DocEng'16)*. 59–68. DOI: 10.1145/2960811.2960817

[61] Taiseer Abdalla Elfadil Eisa, Naomie Salim, and Salha Alzahrani. 2015. Existing plagiarism detection techniques: A systematic mapping of the scholarly literature. *Online Inf. Rev.* 39, 3 (2015), 383–400.

[62] El-Sayed M. El-Alfy, Radwan E. Abdel-Aal, Wasfi G. Al-Khatib, and Faisal Alvi. 2015. Boosting paraphrase detection through textual similarity metrics with abductive networks. *Appl. Soft Comput.* 26, (2015), 444–453. DOI: 10.1016/j.asoc.2014.10.021

[63] Victoria Elizalde. 2013. Using statistic and semantic analysis to detect plagiarism—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[64] Victoria Elizalde. 2014. Using noun phrases and tf-idf for plagiarized document retrieval—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[65] Erik von Elm, Greta Poglia, Bernhard Walder, and Martin R. Tramèr. 2004. Different patterns of duplicate publication: An Analysis of articles used in systematic reviews. *JAMA* 291, 8 (2004), 974–980. DOI: 10.1001/jama.291.8.974

[66] Fezeh Esteki and Faramarz Safi Esfahani. 2016. A plagiarism detection approach based on SVM for persian texts. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'16)*. 149–153.
[67] Asli Eyecioglu and Bill Keller. 2015. Twitter paraphrase identification with simple overlap features and SVMs. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 64–69.
[68] Jody Condit Fagan. 2017. An evidence-based review of academic web search engines, 2014–2016: Implications for librarians' practice and research agenda. *Inf. Technol. Libr.* 36, 2 (2017), 7.
[69] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
[70] Vanessa Wei Feng and Graeme Hirst. 2013. Authorship verification with entity coherence and other rich linguistic features—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.
[71] Rafael Ferreira, George D. C. Cavalcanti, Fred Freitas, Rafael Dueire Lins, Steven J. Simske, and Marcelo Riss. 2018. Combining sentence similarities measures to identify paraphrases. *Comput. Speech Lang.* 47 (2018), 59–73. DOI: 10.1016/j.csl.2017.07.002
[72] Jérémy Ferrero, Frederic Agnes, Laurent Besacier, and Didier Schwab. 2017. CompiLIG at SemEval-2017 Task 1: Cross-language plagiarism detection methods for semantic textual similarity. *arXiv:1704.01346*.
[73] Jérémy Ferrero, Frédéric Agnes, Laurent Besacier, and Didier Schwab. 2017. Using word embedding for cross-language plagiarism detection. *arXiv:1702.03082*.
[74] Jérémy Ferrero, Laurent Besacier, Didier Schwab, and Frédéric Agnes. 2017. Deep investigation of cross-language plagiarism detection methods. *arXiv:1705.08828*.
[75] Tomáš Foltýnek and Irene Glendinning. 2015. Impact of policies for plagiarism in higher education across europe: Results of the project. *Acta Univ. Agric. Silvic. Mendel. Brun.* 63, 1 (2015), 207–216.
[76] Marc Franco-Salvador, Parth Gupta, and Paolo Rosso. 2013. Cross-language plagiarism detection using a multilingual semantic network. In *Advances in Information Retrieval*. 710–713.
[77] Marc Franco-Salvador, Parth Gupta, and Paolo Rosso. 2014. Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. In *Bridging Between Information Retrieval and Databases: PROMISE Winter School 2013*, Nicola Ferro (ed.). Springer-Verlag, Berlin, 227–236. DOI: 10.1007/978-3-642-54798-0_12
[78] Marc Franco-Salvador, Parth Gupta, Paolo Rosso, and Rafael E. Banchs. 2016. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowl.-Based Syst.* 111 (2016), 87–99. DOI: 10.1016/j.knosys.2016.08.004
[79] Marc Franco-Salvador, Paolo Rosso, and Manuel Montes-y-Gómez. 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Inf. Process. Manag.* 52, 4 (2016), 550–570. DOI: 10.1016/j.ipm.2015.12.004
[80] Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 414–423.
[81] Jordan Fréry, Christine Largeron, and Mihaela Juganaru-Mathieu. 2014. UJM at CLEF in Author Identification—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.
[82] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07)*. 1606–1611.
[83] Jean-Gabriel Ganascia, Peirre Glaudes, and Andrea Del Lungo. 2014. Automatic detection of reuses and citations in literary texts. *Lit. Linguist. Comput.* 29, 3 (2014), 412–421. DOI: 10.1093/llc/fqu020
[84] Yasmany García-Mondeja, Daniel Castro-Castro, Vania Lavielle-Castro, and Rafael Muñoz. 2017. Discovering author groups using a b-compact graph-based clustering—notebook for PAN at CLEF 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.
[85] Urvashi Garg and Vishal Goyal. 2016. Maulik: A plagiarism detection tool for hindi documents. *Ind. J. Sci. Technol.* 9, 12 (2016).
[86] Shahabeddin Geravand and Mahmood Ahmadi. 2014. An efficient and scalable plagiarism checking system using bloom filters. *Comput. Electr. Eng.* 40, 6 (2014), 1789–1800.
[87] M. R. Ghaeini. 2013. Intrinsic author identification using modified weighted KNN—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.
[88] Erfaneh Gharavi, Kayvan Bijari, Kiarash Zahirnia, and Hadi Veisi. 2016. A deep learning approach to persian plagiarism detection. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'16)*. 154–159.
[89] Lee Gillam. 2013. Guess again and see if they line up: Surrey's runs at plagiarism detection—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[90] Bela Gipp. 2014. *Citation-based Plagiarism Detection -Detecting Disguised and Cross-language Plagiarism Using Citation Pattern Analysis*. Springer Vieweg Research. Retrieved from http://www.springer.com/978-3-658-06393-1.

[91] Bela Gipp and Norman Meuschke. 2011. Citation pattern matching algorithms for citation-based plagiarism detection: Greedy citation tiling, citation chunking and longest common citation sequence. In *Proceedings of the 11th ACM Symposium on Document Engineering*. 249–258. DOI: 10.1145/2034691.2034741

[92] Bela Gipp, Norman Meuschke, and Joeran Beel. 2011. Comparative evaluation of text- and citation-based plagiarism detection approaches using guttenplag. In *Proceedings of 11th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'11)*. 255–258. DOI: 10.1145/1998076.1998124

[93] Bela Gipp, Norman Meuschke, and Corinna Breitinger. 2014. Citation-based plagiarism detection: Practicability on a large-scale scientific corpus. *J. Assoc. Inf. Sci. Technol.* 65, 8 (2014), 1527–1540. DOI: 10.1002/asi.23228

[94] Bela Gipp, Norman Meuschke, Corinna Breitinger, Jim Pitman, and Andreas Nürnberger. 2014. Web-based demonstration of semantic similarity detection using citation pattern visualization for a cross language plagiarism case. In *Proceedings of the International Conference on Enterprise Information Systems (ICEIS'14)*. 677–683. DOI: 10.5220/0004985406770683

[95] Goran Glavaš, Marc Franco-Salvador, Simone P. Ponzetto, and Paolo Rosso. 2018. A resource-light method for cross-lingual semantic textual similarity. *Knowl.-Based Syst.* 143 (2018), 1–9. DOI: 10.1016/j.knosys.2017.11.041

[96] Lila Gleitman and Anna Papafragou. 2005. Language and thought. In *The Cambridge Handbook of Thinking and Reasoning*, Keith J. Holyoak and Robert G. Morrison (eds.). Cambridge University Press, 633–661.

[97] Demetrios G. Glinos. 2014. A hybrid architecture for plagiarism detection—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*. 958–965.

[98] Helena Gómez-Adorno, Yuridiana Alemán, Darnes Vilariño Ayala, Miguel A Sanchez-Perez, David Pinto, and Grigori Sidorov. 2017. Author clustering using hierarchical clustering analysis—Notebook for PAN at CLEF 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.

[99] Helena Gómez-Adorno, Grigori Sidorov, David Pinto, and Ilia Markov. 2015. A graph based authorship identification approach—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[100] Philipp Gross and Pashutan Modaresi. 2014. Plagiarism alignment detection by merging context seeds—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[101] Deepa Gupta, Vani Kanjirangat, and L. M. Leema. 2016. Plagiarism detection in text documents using sentence bounded stop word n-grams. *J. Eng. Sci. Technol.* 11, 10 (2016), 1403–1420.

[102] Deepa Gupta, Vani Kanjirangat, and Charan Kamal Singh. 2014. Using natural language processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection. In *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI'14)*. 2694–2699. DOI: 10.1109/ICACCI.2014.6968314

[103] Josue Gutierrez, Jose Casillas, Paola Ledesma, Gibran Fuentes, and Ivan Meza. 2015. Homotopy based classification for author verification task—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[104] Yaakov HaCohen-Kerner and Aharon Tayeb. 2017. Rapid detection of similar peer-reviewed scientific papers via constant number of randomized fingerprints. *Inf. Process. Manag.* 53, 1 (2017), 70–86. DOI: 10.1016/j.ipm.2016.06.007

[105] Matthias Hagen, Martin Potthast, and Benno Stein. 2015. Source retrieval for plagiarism detection from large web corpora: Recent approaches. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[106] Osama Haggag and Samhaa Smhaa El-Beltagy. 2013. Plagiarism candidate retrieval using selective query formulation and discriminative query scoring. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[107] Oren Halvani and Lukas Graner. 2017. Author clustering based on compression-based dissimilarity scores—notebook for PAN at CLEF 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.

[108] Oren Halvani and Martin Steinebach. 2014. VEBAV - A simple, scalable and fast authorship verification scheme—notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[109] Oren Halvani, Martin Steinebach, and Ralf Zimmermann. 2013. Authorship verification via k-nearest neighbor estimation—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[110] Oren Halvani and Christian Winter. 2015. A generic authorship verification scheme based on equal error rates—notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[111] Christian Hänig, Robert Remus, and Xose De La Puente. 2015. Exb themis: Extensive feature extraction from word alignments for semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 264–268.

[112] Sarah Harvey. 2014. Author verification using PPM with parts of speech tagging—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[113] Hua He, John Wieting, Kevin Gimpel, Jinfeng Rao, and Jimmy Lin. 2016. UMD-TTIC-UW at SemEval-2016 Task 1: Attention-based multi-perspective convolutional neural networks for textual similarity measurement. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 1103–1108.

[114] Oumaima Hourrane and El Habib Benlahmar. 2017. Survey of plagiarism detection approaches and big data techniques related to plagiarism candidate retrieval. In *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications (BDCA'17)*. 15:1–15:6. DOI: 10.1145/3090354.3090369

[115] Manuela Hürlimann, Benno Weck, Esther van denBerg, Simon Šuster, and Malvina Nissim. 2015. GLAD: Groningen lightweight authorship detection—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[116] Syed Fawad Hussain and Asif Suryani. 2015. On retrieving intelligently plagiarized documents using semantic similarity. *Eng. Appl. Artif. Intell.* 45 (2015), 246–258. DOI: 10.1016/j.engappai.2015.07.011

[117] Ashraf S. Hussein. 2015. A plagiarism detection system for arabic documents. In *Intelligent Systems 2014*, D. Filev, J. Jabłkowski, J. Kacprzyk, M. Krawczak, I. Popchev, L. Rutkowski, V. Sgurev, E. Sotirova, P. Szynkarczyk, and S. Zadrozny (Eds.). Springer International Publishing, 541–552.

[118] Ashraf S. Hussein. 2015. Arabic document similarity analysis using n-grams and singular value decomposition. In *Proceedings of the 2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS'15)*. 445–455. DOI: 10.1109/RCIS.2015.7128906

[119] Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2014. Can characters reveal your native language? A language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1363–1373.

[120] Hideo Itoh. 2016. RICOH at SemEval-2016 Task 1: IR-based semantic textual similarity estimation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 691–695.

[121] Magdalena Jankowska, Vlado Kešelj, and and Evangelos Milios. 2013. Proximity based one-class classification with common n-gram dissimilarity for authorship verification task—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[122] Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios. 2014. Ensembles of proximity-based one-class classifiers for author verification—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[123] Arun Jayapal and Binayak Goswami. 2013. Vector space model and overlap metric for author identification—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[124] Zhuoren Jiang, Miao Chen, and Xiaozhong Liu. 2014. Semantic annotation with rescoredESA: Rescoring concept features generated from explicit semantic analysis. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR'14)*. 25–27. DOI: 10.1145/2663712.2666192

[125] M. A. C. Jiffriya, M. A. C. Akmal Jahan, and Roshan G. Ragel. 2014. Plagiarism detection on electronic text based assignments using vector space model. In *Proceedings of the 7th International Conference on Information and Automation for Sustainability*. 1–5. DOI: 10.1109/ICIAFS.2014.7069593

[126] M. A. C. Jiffriya, M. A. C. Akmal Jahan, Roshan G. Ragel, and Sampath Deegalla. 2013. AntiPlag: Plagiarism detection on electronic submissions of text based assignments. In *Proceedings of the 2013 IEEE 8th International Conference on Industrial and Information Systems*. 376–380. DOI: 10.1109/ICIInfS.2013.6732013

[127] Patrick Juola. 2017. Detecting contract cheating via stylometric methods. In *Proceedings on the Conference on Plagiarism across Europe and Beyond*. 187–198. Retrieved from https://plagiarism.pefka.mendelu.cz/files/proceedings17.pdf.

[128] Patrick Juola and Efstathios Stamatatos. 2013. Overview of the author identification task at PAN 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[129] Rune Borge Kalleberg. 2015. Towards detecting textual plagiarism using machine learning methods. University of Agder. Retrieved from https://brage.bibsys.no/xmlui/bitstream/handle/11250/299460/Rune Borge Kalleberg.pdf?sequence=1.

[130] Rafael-Michael Karampatsis. 2015. CDTDS: Predicting paraphrases in twitter via support vector regression. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 75–79.

[131] Daniel Karaś, Martyna Śpiewak, and Piotr Sobecki. 2017. OPI-JSA at CLEF 2017: Author clustering and style breach detection—Notebook for PAN at CLEF 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.

[132] Roman Kern. 2013. Grammar checker features for author identification and author profiling—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[133] Imtiaz H. Khan, Muazzam A. Siddiqui, Kamal M. Jambi, Muhammad Imran, and Abobakr A. Bagais. 2014. Query optimization in Arabic plagiarism detection: An empirical study. *Int. J. Intell. Syst. Appl.* 7, 1 (2014), 73.

[134] Jamal Ahmad Khan. 2017. Style breach detection: An unsupervised detection model—Notebook for PAN at CLEF 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.

[135] Mahmoud Khonji and Youssef Iraqi. 2014. A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF)—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[136] Khadijeh Khoshnavataher, Vahid Zarrabi, Salar Mohtaj, and Habibollah Asghari. 2015. Developing monolingual persian corpus for extrinsic plagiarism detection using artificial obfuscation—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[137] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. Keele University Technical Report TR/SE-0401. Keele University. 33.

[138] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering—A systematic literature review. *Inf. Softw. Technol.* 51, 1 (2009), 7–15. DOI: 10.1016/j.infsof.2008.09.009

[139] Mirco Kocher. 2016. UniNE at CLEF 2016: Author clustering—Notebook for PAN at CLEF 2016. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'16)*.

[140] Mirco Kocher and Jacques Savoy. 2015. UniNE at CLEF 2015: Author identification—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[141] Mirco Kocher and Jacques Savoy. 2017. UniNE at CLEF 2017: Author clustering—Notebook for PAN at CLEF 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.

[142] Leilei Kong, Yong Han, Zhongyuan Han, Haihao Yu, Qibo Wang, Tinglei Zhang, and Haoliang Qi. 2014. Source retrieval based on learning to rank and text alignment based on plagiarism type recognition for plagiarism detection—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[143] Leilei Kong, Zhimao Lu, Yong Han, Haoliang Qi, Zhongyuan Han, Qibo Wang, Zhenyuan Hao, and Jing Zhang. 2015. Source retrieval and text alignment corpus construction for plagiarism detection—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[144] Leilei Kong, Zhimao Lu, Haoliang Qi, and Zhongyuan Han. 2014. Detecting high obfuscation plagiarism: Exploring multi-features fusion via machine learning. *Int. J. u-and e-Serv. Sci. Technol.* 7, 4 (2014), 385–396.

[145] Leilei Kong, Haoliang Qi, Cuixia Du, Mingxing Wang, and Zhongyuan Han. 2013. Approaches for source retrieval and text alignment of plagiarism detection—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[146] Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *J. Assoc. Inf. Sci. Technol.* 65, 1 (2014), 178–187.

[147] Niraj Kumar. 2014. A graph based automatic plagiarism detection technique to handle artificial word reordering and paraphrasing. In *Computational Linguistics and Intelligent Text Processing*. 481–494.

[148] Marcin Kuta and Jacek Kitowski. 2014. Optimisation of character n-gram profiles method for intrinsic plagiarism detection. In *Artificial Intelligence and Soft Computing*. 500–511.

[149] Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, and Vadim Strijov. 2016. Methods for intrinsic plagiarism detection and author diarization. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'16)*. 912–919. Retrieved from http://ceur-ws.org/Vol-1609/.

[150] Robert Layton, Paul Watters, and Richard Dazeley. 2013. Local n-grams for author identification—notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[151] Paola Ledesma, Gibran Fuentes, Gabriela Jasso, Angel Toledo, and and Ivan Meza. 2013. Distance learning for author verification—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[152] Taemin Lee, Jeongmin Chae, Kinam Park, and Soonyoung Jung. 2013. CopyCaptor: Plagiarized source retrieval system using global word frequency and local feedback—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[153] Chi-kiu Lo, Cyril Goutte, and Michel Simard. 2016. CNRC at SemEeval-2016 task 1: Experiments in crosslingual semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 668–673.

[154] Tara C. Long, Mounir Errami, Angela C. George, Zhaohui Sun, and Harold R. Garner. 2009. Responding to possible plagiarism. *Science* 323, 5919 (2009), 1293–1294. DOI: 10.1126/science.1167408

[155] Ahmed Magooda, Ashraf Y. Mahgoub, Mohsen Rashwan, Magda B. Fayek, and Hazem Raafat. 2015. RDI System for extrinsic plagiarism detection (RDI_RED)—Working Notes for PAN-AraPlagDet at FIRE 2015. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'15)*.

[156] Peyman Mahdavi, Zahra Siadati, and Farzin Yaghmaee. 2014. Automatic external persian plagiarism detection using vector space model. In *Proceedings of the 2014 4th International eConference on Computer and Knowledge Engineering (ICCKE'14)*. 697–702.

[157] Ashraf Y. Mahgoub, Ahmed Magooda, Mohsen Rashwan, Magda B. Fayek, and Hazem Raafat. 2015. RDI System for intrinsic plagiarism detection (RDI_RID)—Working Notes for PAN-AraPlagDet at FIRE 2015. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'15)*.

[158] Promita Maitra, Souvick Ghosh, and Dipankar Das. 2015. Authorship verification - an approach based on random forest—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[159] Cristhian Mayor, Josue Gutierrez, Angel Toledo, Rodrigo Martinez, Paola Ledesma, Gibran Fuentes, and and Ivan Meza. 2014. A single author style representation for the author verification task—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[160] Norman Meuschke and Bela Gipp. 2013. State-of-the-art in detecting academic plagiarism. *Int. J. Educ. Integr.* 9, 1 (2013), 50–71.

[161] Norman Meuschke and Bela Gipp. 2014. Reducing computational effort for plagiarism detection by using citation characteristics to limit retrieval space. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. 197–200.

[162] Norman Meuschke, Christopher Gondek, Daniel Seebacher, Corinna Breitinger, Daniel A. Keim, and Bela Gipp. 2018. An adaptive image-based plagiarism detection approach. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'18)*. DOI: 10.1145/3197026.3197042

[163] Norman Meuschke, Moritz Schubotz, Felix Hamborg, Tomáš Skopal, and Bela Gipp. 2017. Analyzing mathematical content to detect academic plagiarism. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM'17)*. 2211–2214. DOI: 10.1145/3132847.3133144

[164] Norman Meuschke, Nicolas Siebeck, Moritz Schubotz, and Bela Gipp. 2017. Analyzing semantic concept patterns to detect academic plagiarism. In *Proceedings of the 6th International Workshop on Mining Scientific Publications (WOSP'17)*. 46–53. DOI: 10.1145/3127526.3127535

[165] Norman Meuschke, Vincent Stange, Moritz Schubotz, Michael Kramer, and Bela Gipp. 2019. Improving academic plagiarism detection for STEM documents by analyzing mathematical content and citations. In *Proceeedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'19)*.

[166] Pashutan Modaresi and Philipp Gross. 2014. A language independent author verifier using fuzzy c-means clustering—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[167] H. F. Moed, W. J. M. Burger, J. G. Frankfort, and A. F. J. Van Raan. 1985. The application of bibliometric indicators: Important field- and time-dependent factors to be considered. *Scientometrics* 8, 3–4 (1985), 177–203. DOI: 10.1007/BF02016935

[168] Majid Mohebbi and Alireza Talebpour. 2016. Texts semantic similarity detection based graph approach. *Int. Arab J. Inf. Technol.* 13, 2 (2016), 246–251.

[169] Mozhgan Momtaz, Kayvan Bijari, Mostafa Salehi, and Hadi Veisi. 2016. Graph-based approach to text alignment for plagiarism detection in persian documents. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'16)*. 176–179.

[170] Erwan Moreau, Arun Jayapal, and Carl Vogel. 2014. Author verification: exploring a large set of parameters using a genetic algorithm—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[171] Erwan Moreau, Arun Jayapal, Gerard Lynch, and Carl Vogel. 2015. Author verification: Basic stacked generalization applied to predictions from a set of heterogeneous learners—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[172] Erwan Moreau and Carl Vogel. 2013. Style-based distance features for author verification—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[173] Maxim Mozgovoy, Tuomo Kakkonen, and Georgina Cosma. 2010. Automatic student plagiarism detection: Future perspectives. *J. Educ. Comput. Res.* 43, 4 (2010), 511–531.

[174] Aibek Musaev, De Wang, Saajan Shridhar, and Calton Pu. 2015. Fast text classification using randomized explicit semantic analysis. In *Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration*. 364–371. DOI: 10.1109/IRI.2015.62

[175] El Moatez Billah Nagoudi, Ahmed Khorsi, Hadda Cherroun, and Didier Schwab. 2018. 2L-APD: A Two-Level Plagiarism Detection System for Arabic Documents. *Cybern. Inf. Technol.* 18, 1 (2018), 124–138. DOI: 10.2478/cait-2018-0011

[176] Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Clough. 2017. An IR-based approach utilizing query expansion for plagiarism detection in MEDLINE. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 14, 4 (2017), 796–804. DOI: 10.1109/TCBB.2016.2542803

[177] Philip M. Newton. 2018. How common is commercial contract cheating in higher education and is it increasing? A Systematic Review. *Front. Educ.* 3 (2018). DOI: 10.3389/feduc.2018.00067

[178] Le Thanh Nguyen, Nguyen Xuan Toan, and Dinh Dien. 2016. Vietnamese plagiarism detection method. In *Proceedings of the 7th Symposium on Information and Communication Technology (SoICT'16)*. 44–51. DOI: 10.1145/3011077.3011109

[179] Gabriel Oberreuter and Juan D. VeláSquez. 2013. Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Exp. Syst. Appl.* 40, 9 (2013), 3756–3763.

[180] Milan Ojsteršek, Janez Brezovnik, Mojca Kotar, Marko Ferme, Goran Hrovat, Albin Bregant, and Mladen Borovič. 2014. Establishing of a slovenian open access infrastructure: A technical point of view. *Program* 48, 4 (2014), 394–412. DOI: 10.1108/PROG-02-2014-0005

[181] Adeva Oktoveri, Agung Toto Wibowo, and Ari Moesriami Barmawi. 2014. Non-relevant document reduction in anti-plagiarism using asymmetric similarity and AVL tree index. In *Proceedings of the 2014 5th International Conference on Intelligent and Advanced Systems (ICIAS'14)*. 1–5. DOI: 10.1109/ICIAS.2014.6869547

[182] Ahmed Hamza Osman and Naomie Salim. 2013. An improved semantic plagiarism detection scheme based on Chi-squared automatic interaction detection. In *Proceedings of the 2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE'13)*. 640–647. DOI: 10.1109/ICCEEE.2013.6634015

[183] Caleb Owens and Fiona A. White. 2013. A 5-year systematic strategy to reduce plagiarism among first-year psychology university students. *Aust. J. Psychol.* 65, 1 (2013), 14–21. DOI: 10.1111/ajpy.12005

[184] María Leonor Pacheco, Kelwin Fernandes, and Aldo Porco. 2015. Random forest with increased generalization: A universal background approach for authorship verification—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[185] Yurii Palkovskii and Alexei Belov. 2013. Using hybrid similarity methods for plagiarism detection—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[186] Yurii Palkovskii and Alexei Belov. 2014. Developing high-resolution universal multi-type n-gram plagiarism detector. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*. 984–989.

[187] Guy Paré, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. 2015. Synthesizing information systems knowledge: A typology of literature reviews. *Inf. Manag.* 52, 2 (2015), 183–199. DOI: 10.1016/j.im.2014.08.008

[188] Merin Paul and Sangeetha Jamal. 2015. An Improved SRL based plagiarism detection technique using sentence ranking. *Procedia Comput. Sci.* 46 (2015), 223–230. DOI: 10.1016/j.procs.2015.02.015

[189] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 425–430.

[190] Jian Peng, Kim-Kwang Raymond Choo, and Helen Ashman. 2016. Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *J. Netw. Comput. Appl.* 70 (2016), 171–182. DOI: 10.1016/j.jnca.2016.04.001

[191] Solange de L. Pertile, Viviane P. Moreira, and Paolo Rosso. 2015. Comparing and combining content- and citation-based approaches for plagiarism detection. *J. Assoc. Inf. Sci. Technol.* 67, 10 (2015), 2511–2526. DOI: 10.1002/asi.23593

[192] Solange de L. Pertile, Paolo Rosso, and Viviane P. Moreira. 2013. Counting co-occurrences in citations to identify plagiarised text fragments. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. 150–154.

[193] Timo Petmanson. 2013. Authorship identification using correlations of frequent features—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[194] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1341–1351.

[195] Gaspar Pizarro V. and Juan D. Velásquez. 2017. Docode 5: Building a real-world plagiarism detection system. *Eng. Appl. Artif. Intell.* 64 (Jun. 2017), 261–271. DOI: 10.1016/j.engappai.2017.06.001

[196] Juan-Pablo Posadas-Durán, Grigori Sidorov, Ildar Batyrshin, and Elibeth Mirasol-Meléndez. 2015. Author verification using syntactic n-grams—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[197] Martin Potthast, Tim Gollub, Matthias Hagen, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th International Competition on Plagiarism Detection. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[198] Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein. 2014. Overview of the 6th International Competition on Plagiarism Detection. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[199] Martin Potthast, Matthias Hagen, and Benno Stein. 2016. Author Obfuscation: Attacking the state of the art in authorship verification. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'16)*.

[200] Martin Potthast, Francisco Rangel, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. 2017. Overview of PAN'17: Author identification, author profiling, and author obfuscation. In *Proceedings of the 7th International Conference of the CLEF Initiative*. DOI: 10.1007/978-3-319-65813-1_25

[201] Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING'10)*. 997–1005.

[202] Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. 2009. Overview of the 1st international competition on plagiarism detection. In *Proceedings of the SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN'09)*. 1–9.

[203] Amit Prakash and Sujan Kumar Saha. 2014. Experiments on document chunking and query formation for plagiarism source retrieval—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[204] Piotr Przybyła, Nhung T. H. Nguyen, Matthew Shardlow, Georgios Kontonatsios, and Sophia Ananiadou. 2016. NaCTeM at SemEval-2016 Task 1: Inferring sentence-level semantic similarity from an ensemble of complementary lexical and sentence-level features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 614–620.

[205] Javad Rafiei, Salar Mohtaj, Vahid Zarrabi, and Habibollah Asghari. 2015. Source retrieval plagiarism detection based on noun phrase and keyword phrase extraction—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[206] Shima Rakian, Esfahani Faramarz Safi, and Hamid Rastegari. 2015. A Persian fuzzy plagiarism detection approach. *J. Inf. Syst. Telecommun.* 3, 3 (2015), 182–190.

[207] N Riya Ravi and Deepa Gupta. 2015. Efficient paragraph based chunking and download filtering for plagiarism source retrieval—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[208] N. Riya Ravi, Vani Kanjirangat, and Deepa Gupta. 2016. Exploration of fuzzy C means clustering algorithm in external plagiarism detection system. In *Intelligent Systems Technologies and Applications*. Springer, 127–138.

[209] Andi Rexha, Stefan Klampfl, Mark Kröll, and Roman Kern. 2015. Towards authorship attribution for bibliometrics using stylometric features. In *Proceedings of the Conference on Computational Linguistics and Bibliometrics co-located with the International Conference on Scientometrics and Informetrics (CLBib@ ISSI)*. 44–49.

[210] Diego Antonio Rodríguez Torrejón and José Manuel Martín Ramos. 2014. CoReMo 2.3 Plagiarism detector text alignment module—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[211] Paolo Rosso, Francisco Rangel, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein. 2016. Overview of PAN'16. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. 332–350.

[212] Frantz Rowe. 2014. What literature review is not: Diversity, boundaries and recommendations. *Eur. J. Inf. Syst.* 23, 3 (2014), 241–255. DOI: 10.1057/ejis.2014.7

[213] Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung Poland NLP Team at SemEval-2016 Task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 602–608.

[214] Kamil Safin and Rita Kuznetsova. 2017. Style breach detection with neural sentence embeddings—Notebook for PAN at CLEF 2017. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.

[215] Anuj Saini and Aayushi Verma. 2016. Anuj@ DPIL-FIRE2016: a novel paraphrase detection method in hindi language using machine learning. In *Proceedings of the Forum for Information Retrieval Evaluation*. 141–152.

[216] Miguel A. Sanchez-Perez, Alexander Gelbukh, and Grigori Sidorov. 2015. Dynamically adjustable approach through obfuscation type recognition—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[217] Miguel A Sanchez-Perez, Grigori Sidorov, and Alexander F Gelbukh. 2014. A winning approach to text alignment for text reuse detection at PAN 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*. 1004–1011.

[218] Fernando Sánchez-Vega, Esaú Villatoro-Tello, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, and Paolo Rosso. 2013. Determining and characterizing the reused text for plagiarism detection. *J. Assoc. Inf. Sci. Technol.* 65, 5 (2013), 1804–1813. DOI: 10.1016/j.eswa.2012.09.021

[219] Yunita Sari and Mark Stevenson. 2015. A machine learning-based intrinsic method for cross-topic and cross-genre authorship verification—Notebook for PAN at CLEF 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[220] Yunita Sari and Mark Stevenson. 2016. Exploring word embeddings and character n-grams for author clustering—Notebook for PAN at CLEF 2016. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'16)*.

[221] Satyam, Anand, Arnav Kumar Dawn, and and Sujan Kumar Saha. 2014. Statistical analysis approach to author identification using latent semantic analysis—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[222] Taneeya Satyapanich, Hang Gao, and Tim Finin. 2015. Ebiquity: Paraphrase and semantic similarity in twitter using skipgrams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 51–55.

[223] Andreas Schmidt, Reinhold Becker, Daniel Kimmig, Robert Senger, and Steffen Scholz. 2014. A concept for plagiarism detection based on compressed bitmaps. In *Procceedings of the 6th International Conference on Advances in Databases, Knowledge, and Data Applications*. 30–34.

[224] Shachar Seidman. 2013. Authorship verification using the impostors method—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF13)*.

[225] Prasha Shrestha, Suraj Maharjan, and Thamar Solorio. 2014. Machine translation evaluation metric for text alignment—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[226] Prasha Shrestha and Thamar Solorio. 2013. Using a variety of n-grams for the detection of different kinds of plagiarism. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[227] Muazzam Ahmed Siddiqui, Imtiaz Hussain Khan, Kamal Mansoor Jambi, Salma Omar Elhaj, and Abobakr Bagais. 2014. Developing an arabic plagiarism detection corpus. *Comput. Sci. Inf. Technol.* 4, 2014 (2014), 261–269. DOI: 10.5121/csit.2014.41221

[228] L. Sindhu and Sumam Mary Idicula. 2015. Fingerprinting based detection system for identifying plagiarism in malayalam text documents. In *Proceedings of the 2015 International Conference on Computing and Network Communications (CoCoNet'15)*. 553–558. DOI: 10.1109/CoCoNet.2015.7411242

[229] Abdul Sittar, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. 2016. Author diarization using cluster-distance approach. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'16)*. 1000–1007.

[230] Sidik Soleman and Ayu Purwarianti. 2014. Experiments on the Indonesian plagiarism detection using latent semantic analysis. In *Proceedings of the 2014 2nd International Conference on Information and Communication Technology (ICoICT'14)*. 413–418. DOI: 10.1109/ICoICT.2014.6914098

[231] Hussein Soori, Michal Prilepok, Jan Platos, Eshetie Berhan, and Vaclav Snasel. 2014. Text similarity based on data compression in Arabic. In *AETA 2013: Recent Advances in Electrical Engineering and Related Sciences*. Springer, 211–220.

[232] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. Overview of the author identification task at PAN 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[233] Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. 2015. Overview of the PAN/CLEF 2015 Evaluation Lab. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the 6th International Conference of the CLEF Initiative (CLEF'15)*. 518–538. DOI: 10.1007/978-3-319-24027-5_49

[234] Efstathios Stamatatos, Walter Daelemans Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. Overview of the author identification task at PAN 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[235] Benno Stein, Sven zu Eissen, and Martin Potthast. 2007. Strategies for retrieving plagiarized documents. In *Proceedings of the 30th Annual International ACM SIGIR Conference*. 825–826. DOI: 10.1145/1277741.1277928

[236] Imam Much Ibnu Subroto and Ali Selamat. 2014. Plagiarism detection through internet using hybrid artificial neural network and support vectors machine. *Telecommun. Comput. Electron. Control.* 12, 1 (2014), 209–218.

[237] Šimon Suchomel and Michal Brandejs. 2014. Heterogeneous queries for synoptic and phrasal search—Notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[238] Šimon Suchomel and Michal Brandejs. 2015. Improving synoptic querying for source retrieval. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

[239] Šimon Suchomel, Jan Kasprzak, and Michal Brandejs. 2013. Diverse queries and feature type selection for plagiarism discovery—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[240] M. D. Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. DLS@CU: Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval'14)*. 241–246.

[241] M. D. Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Trans. Assoc. Comput. Linguist.* 2 (2014), 219–230.

[242] M. D. Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 148–153.

[243] Junfeng Tian and Man Lan. 2016. ECNU at SemEval-2016 Task 1: Leveraging word embedding from macro and micro views to boost performance for semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 621–627.

[244] Diego A. Rodríguez Torrejón and José Manuel Martín Ramos. 2013. Text alignment module in CoReMo 2.1 plagiarism detector. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[245] Michael Tschuggnall and Günther Specht. 2013. Detecting plagiarism in text documents through grammar-analysis of authors. *Datenbanksysteme für Business, Technologie und Web (BTW) 2028*, Volker Markl, Gunter Saake, Kai-Uwe Sattler, Gregor Hackenbroich, Bernhard Mitschang, Theo Härder, and Veit Köppen (Eds.). Gesellschaft für Informatik e.V., 241–259.

[246] Michael Tschuggnall and Günther Specht. 2013. Using grammar-profiles to intrinsically expose plagiarism in text documents. In *Natural Language Processing and Information Systems*. 297–302.

[247] Michael Tschuggnall, Efstathios Stamatatos, Ben Verhoeven, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2017. Overview of the author identification task at PAN-2017: Style breach detection and author clustering. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17)*.

[248] Alper Kursat Uysal and Serkan Gunal. 2014. Text classification using genetic algorithm oriented latent semantic features. *Exp. Syst. Appl.* 41, 13 (2014), 5938–5947. DOI: 10.1016/j.eswa.2014.03.041

[249] Vani Kanjirangat and Deepa Gupta. 2014. Using K-means cluster based techniques in external plagiarism detection. In *Proceedings of the 2014 International Conference on Contemporary Computing and Informatics (IC3I'14)*. 1268–1273. DOI: 10.1109/IC3I.2014.7019659

[250] Vani Kanjirangat and Deepa Gupta. 2015. Investigating the impact of combined similarity metrics and POS tagging in extrinsic text plagiarism detection system. In *Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI'15)*. 1578–1584. DOI: 10.1109/ICACCI.2015.7275838

[251] Vani Kanjirangat and Deepa Gupta. 2016. Study on extrinsic text plagiarism detection techniques and tools. *J. Eng. Sci. Technol. Rev.* 9, 5 (2016), 9–23.

[252] Vani Kanjirangat and Deepa Gupta. 2017. Detection of idea plagiarism using syntax–semantic concept extractions with genetic algorithm. *Exp. Syst. Appl.* 73 (2017), 11–26. DOI: 10.1016/j.eswa.2016.12.022

[253] Vani Kanjirangat and Deepa Gupta. 2017. Identifying document-level text plagiarism: A two-phase approach. *J. Eng. Sci. Technol.* 12, 12 (2017), 3226–3250.

[254] Vani Kanjirangat and Deepa Gupta. 2017. Text plagiarism classification using syntax based linguistic features. *Exp. Syst. Appl.* 88 (2017), 448–464. DOI: 10.1016/j.eswa.2017.07.006

[255] Anna Vartapetiance and Lee Gillam. 2013. A textual modus operandi: surrey's simple system for author identification—notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[256] Juan D. Velásquez, Yerko Covacevich, Francisco Molina, Edison Marrese-Taylor, Cristián Rodríguez, and Felipe Bravo-Marquez. 2016. DOCODE 3.0 (DOcument COpy DEtector): A system for plagiarism detection by applying an information fusion process from multiple documental data sources. *Inf. Fus.* 27 (2016), 64–75. DOI: 10.1016/j.inffus.2015.05.006

[257] Ondřej Veselý, Tomáš Foltýnek, and Jiří Rybička. 2013. Source retrieval via naïve approach and passage selection heuristics—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[258] Darnes Vilariño, David Pinto, Helena Gómez, Saúl León, and Esteban Castillo. 2013. Lexical-syntactic and graph-based features for authorship verification—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[259] Ngoc Phuoc An Vo, Octavian Popescu, and Tommaso Caselli. 2014. FBK-TR: SVM for semantic relatedeness and corpus patterns for RTE. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval'14)*. 289–293.

[260] Hai Hieu Vu, Jeanne Villaneau, Farida Saïd, and Pierre-François Marteau. 2014. Sentence similarity by combining explicit semantic analysis and overlapping N-grams. In *Text, Speech and Dialogue*. 201–208.

[261] Elizabeth Wager. 2014. Defining and responding to plagiarism. *Learn. Publ.* 27, 1 (2014), 33–42. DOI: 10.1087/20140105

[262] Wafa Wali, Bilel Gargouri, and Abdelmajid Ben Hamadou. 2015. Supervised learning to measure the semantic similarity between arabic sentences. In *Computational Collective Intelligence*. 158–167.

[263] John Walker. 1998. Student plagiarism in universities: What are we doing about it? *High. Educ. Res. Dev.* 17, 1 (1998), 89–106. DOI: 10.1080/0729436980170105

[264] Shuai Wang, Haoliang Qi, Leilei Kong, and Cuixia Nu. 2013. Combination of VSM and jaccard coefficient for external plagiarism detection. In *Proceedings of the 2013 International Conference on Machine Learning and Cybernetics*. 1880–1885. DOI: 10.1109/ICMLC.2013.6890902

[265] Debora Weber-Wulff. 2014. *False feathers: A Perspective on Academic Plagiarism*. Springer, Berlin.

[266] Debora Weber-Wulff, Christopher Möller, Jannis Touras, and Elin Zincke. 2013. *Plagiarism Detection Software Test* 2013. Retrieved from http://plagiat.htw-berlin.de/wp-content/uploads/Testbericht-2013-color.pdf.

[267] Agung Toto Wibowo, Kadek W. Sudarmadi, and Ari M. Barmawi. 2013. Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on Bahasa Indonesia documents. In *Proceedings of the 2013 International Conference of Information and Communication Technology (ICoICT'13)*. 128–133. DOI: 10.1109/ICoICT.2013.6574560

[268] Kyle Williams, Hung-Hsuan Chen, Sagnik Ray Chowdhury, and C. Lee Giles. 2013. Unsupervised ranking for plagiarism source retrieval—Notebook for PAN at CLEF 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'13)*.

[269] Kyle Williams, Hung-Hsuan Chen, and C. Lee Giles. 2014. Classifying and ranking search engine results as potential sources of plagiarism. In *Proceedings of the 2014 ACM Symposium on Document Engineering (DocEng'14)*. 97–106. DOI: 10.1145/2644866.2644879

[270] Kyle Williams, Hung-Hsuan Chen, and C. Lee Giles. 2014. Supervised ranking for plagiarism source retrieval—notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[271] Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. A lightweight and high performance monolingual word aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 702–707.

[272] Takeru Yokoi. 2015. Sentence-based plagiarism detection for japanese document based on common nouns and part-of-speech structure. In *Intelligent Software Methodologies, Tools and Techniques*. 297–308.

[273] Guido Zarrella, John Henderson, Elizabeth M. Merkhofer, and Laura Strickhart. 2015. MITRE: Seven systems for semantic similarity in tweets. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 12–17.

[274] Chunxia Zhang, Xindong Wu, Zhendong Niu, and Wei Ding. 2014. Authorship identification from unstructured texts. *Knowl.-Based Syst.* 66 (2014), 99–111. DOI: 10.1016/j.knosys.2014.04.025

[275] Jiang Zhao and Man Lan. 2015. Ecnu: Leveraging word embeddings to boost performance for paraphrase in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 34–39.

[276] Valentin Zmiycharov, Dimitar Alexandrov, Hristo Georgiev, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2016. Experiments in authorship-link ranking and complete author clustering—Notebook for PAN at CLEF 2016. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'16)*.

[277] Sven Meyer Zu Eissen and Benno Stein. 2006. Intrinsic plagiarism detection. In *Proceedings of the European Conference on Information Retrieval*. 565–569.

[278] Denis Zubarev and Ilya Sochenkov. 2014. Using sentence similarity measure for plagiarism source retrieval—notebook for PAN at CLEF 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.

[279] Teddi Fishman. 2009. We know it when we see it' is not good enough: Toward a standard definition of plagiarism that transcends theft, fraud, and copyright. In *Proceedings 4th Asia Pacific Conference on Educational Integrity (4APCEI'09)*. 5. https://www.bmartin.cc/pubs/09-4apcei/4apcei-Fishman.pdf.

Received March 2019; revised August 2019; accepted August 2019