International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 2
ISSN: 2088-5334

Cross-Language Plagiarism Detection: Methods, Tools, and Challenges: A Systematic Review

Miguel Botto-Tobar [a,b,*], Alexander Serebrenik [a], Mark G.J. van den Brand [a]

[a] *Eindhoven University of Technology, The Netherlands*
[b] *Research Group in Artificial Intelligence and Information Technology, University of Guayaquil, Ecuador*
Corresponding author: [*]m.a.botto.tobar@tue.nl

***Abstract*— Plagiarism is one of the most serious academic offenses. However, people have adopted different approaches to avoid plagiarism, such as transcribing excerpts from one language. Thus, it is challenging to realize this plagiarism form unless someone fully understands another language. Researchers have developed approaches for detecting plagiarism in a variety of different languages. However, most methods created in the past have proved effective for detecting plagiarism in papers published in a single language, most notably English. Therefore, this paper aims to provide a systematic literature review of cross-language plagiarism detection methods (CLPD) in a natural language context. The approach used to perform this study consisted of an extensive search for relevant literature through an SLR and Snowballing. Therefore, we present an overview of (i) cross-language plagiarism detection techniques; (ii) the artifacts and the aspects that were considered in the evaluation phase; and (iii) the lack of guidelines and tools for its implementation. Its contribution lies in its ability to highlight emerging cross-language plagiarism detection techniques trends. Further, we identify any of these techniques in other domains, for instance, software engineering.**

***Keywords*— Cross-language; plagiarism detection; SLR; snowballing.**

Manuscript received 17 Mar. 2021; revised 14 Jun. 2021; accepted 7 Mar. 2022. Date of publication 30 Apr. 2022.

I. INTRODUCTION

Nowadays, the web has a considerable amount of information easily accessible by users through the Internet; its easy access has become a concern for scholars due to protected content that other people can reuse or copy without acknowledging the source. In this context, the term "plagiarism" appears, and according to IEEE, plagiarism "is the reuse of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source" [1]. Hence, this problem occurs when someone else shows a work of another, omits the quotation marks when using a quote, or provides false information about the source of a quote without crediting the source. Some studies have been conducted to tackle this problem, demonstrating that the current methods are limited as to the comparison of sources through plagiarism detection systems.

Potthast et al. [2] presented a framework for evaluating plagiarism detection using a corpus, and their investigation demonstrates that the process of developing specialized training corpora for plagiarism detection may be automated and thus carried out on a wide scale. Alzahrani et al. [3] distinguished between literal and intelligent plagiarism from the perspective of the plagiarist's conduct and also discussed systematic frameworks and methods for detecting monolingual, extrinsic, intrinsic, and cross-lingual plagiarism using plagiarism categories. Nevertheless, new challenges are exacerbated today by plagiarism across different languages [3], which means plagiarism by translation (machine translation tools or humans), e.g., a text is translated or reused (totally or partially) from one language to another either in a bit different style or using synonyms/antonyms in order to obfuscate the detection process. Therefore, this kind of plagiarism is more challenging than the other plagiarism categories due to the difficulty of retrieving suspicious documents from a large multilingual corpus [4]. Thus, a variety of methods for cross-language plagiarism detection have been published in the literature. Specifically, Barrón-Cedeño et al. [4] proposed architecture for plagiarism detection across languages: heuristic retrieval, detailed analysis, and post-processing, and explored its suitability through three cross-language similarity models. However, efforts to automatically detect cross-language plagiarism depend on a preliminary translation, which is not always

available. Further, Franco-Salvador et al. [5] examined the contributions of knowledge graphs to cross-language plagiarism detection in three areas: word meaning disambiguation, vocabulary extension, and representation through similarities to a collection of ideas; Ferrero et al. [6] studied cross-language plagiarism detection techniques across six language pairings and two granularities of text units in order to reach solid findings of the best algorithms while also doing in-depth analyses of connections across document types and languages; and Tlitova et al. [7]. reviewed the available techniques for detecting cross-language plagiarism in scientific publications, with a particular emphasis on the Russian-English language pair

These studies attempted to provide quality information that can assist in the detection process [2], [3], [8]–[10]. Despite the authors' best efforts and different evidence-based recommendations for cross-language plagiarism detection, their implementation remains difficult due mostly to the differences in linguistic structures across languages.

The purpose of this study is to discuss cross-language plagiarism detection methods applied in a natural language context, as well as the resulting findings through the following research question: *What Cross-Language Plagiarism Detection (CLPD) methods have been employed in a Natural Language (NL) context by practitioners and researchers, and how were they used?* Since our research question is too broad, it has been decomposed into more detailed sub-questions:

RQ1: What techniques are employed for cross-language plagiarism detection (CLPD)? This question tries to identify the CLPD detection methods. First, it identifies that CLPD is difficult to execute as there are barriers as Barrón-Cedeño et al. [4] explain:

- Many people understand source context better when it is translated into their native language rather than their second;
- Content in a variety of languages is available. Due to the necessity of reaching a large audience, most people have been forced to make their content available in multiple languages.

RQ2: How are the proposed techniques evaluated? The question aims to give insights into the CLPD detection methods identified in RQ1. Therefore, this question will aim to evaluate the practicality of the CLPD techniques, i.e., whether the methods can work or not in various contexts. For instance, how the identified tools behave in varying cases of sentence manipulation.

RQ3: What is the available support for the identified techniques? This question first recognizes that there may be drawbacks to employing the identified CLPD methods due to their complexity and difficulty.

In order to provide a balanced and objective summary of research evidence of CLPD methods, we have chosen to carry out a systematic literature review (SLR) and Snowballing. This study presents an overview of i) Cross-language plagiarism detection techniques; ii) Artifacts used in the evaluation phase; iii) Tools, and iv) Research challenges and future work.

This paper is organized as follows. Section II. describes the protocol followed in carrying out the literature review. Section III presents the results obtained; also, it discusses the findings of this study; and it analyses the threats to the validity of the results. Finally, Section IV presents our conclusions and suggests areas for further investigation.

II. Materials And Method

To perform a literature review, two commonly used techniques are Snowballing [11], [12] and Systematic Literature Review (SLR) [13]. An SLR aims at categorizing and summarizing the existing information about a phenomenon of interest (e.g., a particular research question) in an unbiased manner [13], and its counterpart, snowballing, consists of iteratively following the citations of a small collection of randomly identified papers. However, several core papers might have hundreds of citations, and rendering snowballing might be too labor-intensive. Thus, we applied for a systematic literature review as the research methodology for reviewing the literature, given that it is the most appropriate technique to answer our Research Questions (RQs). It needs a well-defined search procedure and rigorous criteria for filtering and selecting relevant papers.

A. Searching

The target population for this review, and hence the fundamental inclusion criteria, are studies that propose, evaluate, or validate cross-language plagiarism detection techniques. The three main concepts are the difference in language, copy, and detection, and we use them to identify alternative terms and/or synonyms, as shown in TABLE I, for formulating the search string for the database search. Then, we combine them to make a general search string that takes the form of C1 AND C2 AND C3 (78 = 13 X 3 X 2 combinations).

TABLE I
SEARCH STRING APPLIED

	Concept	Alternative terms & Synonyms
C1	difference in language	"cross language" OR "crosslanguage" OR "cross lingual" OR "crosslingual" OR "cross linguistic" OR "crosslinguistic" OR "multi language" OR "multilanguage" OR "multi lingual" OR "multilingual" OR "multi linguistic" OR "multilinguistic" OR "machine translation"
C2	copy	copy OR duplicate OR plagiarism
C3	detection	detection OR discovery

We first searched for primary studies in IEEExplore, the ACM digital libraries, Springer Link, and Science Direct. However, our results contained multiple inconsistencies, e.g., in IEEEXplore, by adding an OR to our query reduced the number of results. Although the search terms were presented in their abstracts, ACM DL and ScienceDirect missed some papers. Hence, we opted for Google Scholar since it provides extensive coverage of different electronic sources [14], and as recommended by Keele University [13] and was conducted by Landman et al. [15].

In February 2019, we performed our automatic search with our search string to retrieve studies and obtained 130K

references (hits). After that, duplicate studies, Google books, and non-English references were eliminated (to avoid downloading non-English PDF files), resulting in 46K references (hits). Next, we proceeded to download all studies (PDF files), and it resulted in 30K PDF files. It should be noted that our tool was not able to download the rest of the references (16K) since i) some papers are behind the paywall, ii) the paper's web page organization has a different formed URL; iii) some papers are not available anymore, e.g., academia.eu. Later, pdf2text was applied to each PDF file, and lastly, we retained only papers having five or more pages and written in English. In this way, we have obtained 27K documents.

Since manual analysis of 27K documents is unfeasible, we followed the approach of Landman et al. [15] to reduce the number of potentially relevant documents. This approach aims at establishing criteria based on the frequency of keywords (Table 1) in the full text, the first 20% of the text (head), and the last 20% excluding the references/bibliography (tail). Hence, thresholds were defined as the number of the frequency of keywords in both head and tail. We validated all thresholds (5 to 1 frequency of keywords) of these criteria by sampling beyond the thresholds and manually scanning the additional papers for false negatives. This process preserved 170 documents (0.001% of the original set).

B. Inclusion and exclusion criteria

We performed an iterative relevance assessment by selecting random samples of 15 papers. The first and second authors of this study read the title, abstract, and conclusion of each paper in order to label those that met any inclusion/exclusion criteria (TABLE II) and, moreover, with the questions as follows:

- The study presents methods and techniques for cross-language plagiarism detection.
- The study presents tools for cross-language plagiarism detection.

TABLE II
INCLUSION AND EXCLUSION CRITERIA

Criteria	AI Assessment criteria
Inclusion	Primary studies.
Inclusion	Studies (papers) that address methodologies, methods or techniques on cross-language plagiarism detection.
Exclusion	Studies written in any language other than the English language.
Exclusion	Short publications and posters (4 pages).

The possible answers to these questions were: I agree (1), and I do not agree (0). At the end of each round, we compute Cohen's k [16] to measure the agreement between the ratters and discuss the disagreements. We repeat the process until k exceeds 0.6. Indeed, $0.61 \leq k \geq 0.80$ should be interpreted as substantial agreement [16], [17]; this range has also been used in previous studies [18]. When the acceptable agreement level had been reached, the first author continued the selection procedure independently, and this step produced 59 documents. Further, the search-based approach is sensitive to the choice of the keywords, so to compensate for this, we have performed the snowballing method [11].

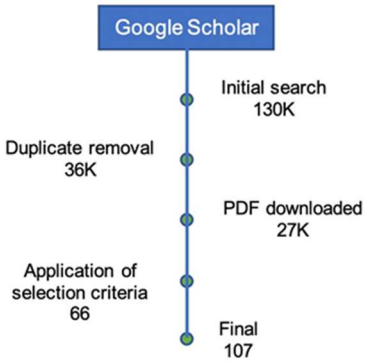

Fig. 1 Search and selection process

To explore the references of the 59 selected papers. In this stage, we only did it once and found 20 potentially related papers; and finally, based on applying inclusion and exclusion criteria, we selected 7 papers from the 20 papers. Thereby, we obtained 66 papers.

On the other hand, due to the high number of references without any of their PDF files (46K references - 30K downloaded = 16K missing), we opted to perform a title-based search, e.g., to select all references that contain at least one keyword (Table 1) in the title. Hence, we got 3,428 references and 2,827 PDF files. We applied 1) the Landman et al. [15] approach and kept 168 papers, and then 2) our inclusion and exclusion criteria over these 168 papers, preserving 66 papers. Consequently, the final total of selected papers for this SLR is 107: 59 (SLR) + 7 (snowballing) + 41 (title-based search). Figure Fig. 1 presents the summary of the process to select the papers.

III. RESULTS AND DISCUSSIONS

A. Review results

1) RQ1. What techniques are employed for cross-language plagiarism detection (CLPD)?

To address RQ1, we extend the early classification of cross-language similarity analysis techniques [19], [20], as follows:

- Translation-based and Monolingual Analysis (T+MA).
- Dictionary and thesaurus-based approaches.
- Parallel corpora-based models.
- Comparable corpora-based models.

TABLE III shows the distribution of selected studies [19], [20]. This indicates that most of the selected studies present a translation-based and monolingual analysis (T+MA) model. Specifically, using Google Translator as a public translation service is frequently employed.

Translation-based and Monolingual Analysis (T+MA). This uses the MT system to translate suspicious texts into the same language as original texts and then a monolingual comparison. Kasprzak & Brandejs presented a method for intrinsic plagiarism detection in the PAN 2010 plagiarism detection competition [21]. They utilized the Czech National Archive of Graduate Theses and many other production systems to create the best performing solution in the PAN 2010 plagiarism detection competition (millions of text documents). However, utilizing public translation services such as Google Translate similar to [3], [22]–[26], or Yahoo!

Babelfish is incompatible with huge collections of documents. Combining an information retrieval approach, Pataki showed a technique for detecting cross-language plagiarism via machine translation [23]. The system could identify a 10-sentence translation with a probability of over 95% for the German-English language pair and 99% for the Hungarian-English language pair when tested on a machine-translated corpus. However, the accuracy measurements did not provide useful results since the database contains an excessive number of duplicate items identified as false positives, compared to the Alzahrani [25] study, which conducted research on the semantic similarities between Arabic and English in short phrases and sentences. From a monolingual viewpoint, they used dictionary and machine translation methods to determine the relatedness of the cross-lingual texts. The authors used a Pearson correlation coefficient to compare the findings to the human evaluations, and they were triangulated with the best, worst, and mean for all human participants. However, further statistical analysis showed no significant difference between both algorithms and the humans' judgment; and interestingly, Safi-Esfahani et al. [26] presented a framework for cross-lingual plagiarism analysis and detection of plagiarism. Their tests show that the enhanced translation tools increase the proposed method's accuracy. On the other hand, Kothwal and Varma [27] identified suspicious documents produced via text reuse of previously published publications across distant language pairings, such as Arabic and Indian. Their method was twofold: (1) they used key phrases rather than n-grams, and (2) they developed a new similarity measure. However, the F-measure yielded: 1) 0.649; and 2) 0.608; and Kent & Salim proposed a web-based approach to detecting cross-language plagiarism by implementing different techniques and tools (Google translate, Google Search API) to assist the detection process, and it also integrated the fingerprint matching technique [24]. However, each K-gram requires K bytes of storage, and hence the space-consuming becomes too large for larger values of K.

Dictionary and thesaurus-based approaches. This approach translates single words or concepts (e.g., locations, dates, numbers, expressions) from language L to language L' using bilingual dictionaries [18], and then performs the plagiarism analysis using such methods as Vector Space Model (CL-VSM) or Conceptual Thesaurus Similarity (CL-CTS). Gupta et al. [28] analyzed the monolingual paraphrases of English and cross-lingual paraphrases of German and Spanish. This work was based on VSM, and they expect the results to be better when a synonym addition strategy is employed using thesauri, dictionary [29], [30], or Wordnet. One of the approaches using WordNet is MLPlag [31]. Nevertheless, incomplete WordNet may cause difficulties, especially when dealing with less common languages [31]. Further, Gupta et al. [32] also developed a concept-based similarity model and tested it using the Eurovoc conceptual thesaurus on three distinct corpora of varying types and two language pairings, English-German and English-Spanish. This approach was very general, provided competitive outcomes, and was extremely stable and constant among corpora. However, it did not make any comparisons to statistical conceptual models.

On the other hand, seeing the impact of available resources like bi-lingual dictionary, Gupta & Singhal used Okapi BM25 model to calculate the similarity between document pairs [29]. Results suggest that available resources can find the text reuse document pairs for Hindi-English. Nonetheless, they need to work on the precision of the system to see how the system performs for different amount and nature of text re-use; and regarding to linear transformations in bi-lingual dictionaries, Brychcín experimented with unsupervised techniques for sentence similarity and showed significantly improved by the word weighting [30].

TABLE III
DISTRIBUTION OF SELECTED STUDIES BASED ON [19], [20] CLASSIFICATION

Model		Frequency	Selected studies
Syntax-based models	CL-CNG	3	[82]–[84]
Dictionary-based models	CL-VSM	5	[28], [29], [58], [77], [85]
	CL-CTS	9	[30], [38], [52], [68], [77], [79], [86]–[88]
Semantic-based models*	CL-WE	17	[89], [43], [51], [53], [62], [65], [69]–[71], [76], [90]–[96]
	CL-LSA	2	[97], [98]
Parallel corpora-based models	CL-ASA	7	[33], [34], [99]–[103]
	CL-LSI	1	[35]
Comparable corpora-based models	CL-KGA	5	[49], [66], [104]–[106]
	CL-ESA	1	[107]
Fuzzy-based models*	CL-FUZZY	5	[75], [80], [108]–[110]
MT-based models	T + MA	27	[23], [24], [25], [26], [83]–[88], [89]–[98], [99]–[105]

** Other models identified in this study.*

Parallel corpora-based models. This model uses documents in different languages which describe the same topic. Then machine learning techniques such as Latent Semantic Indexing (CL-LSI) and Alignment-based Similarity Analysis (CL-ASA) are applied to the aligned corpus. For example, Pinto et al. [33] applied the IBM alignment model 1 in 3-tasks: text classification, information retrieval, and plagiarism analysis to obtain a statistical bilingual dictionary to approximate the relatedness probability of two given documents (written in different languages). The results obtained highlight the benefit of using the presented statistical approach. Further, Yahyaei et al. [34] developed an algorithm to perform cross-lingual text fragment alignment based on models of divergence from randomness. The results showed that a one-stage direct computation of similarity using a probabilistic dictionary (lexical probabilities) significantly outperforms a method that translates and summaries the documents and estimates a monolingual similarity between the documents.

On the other hand, Mostafa and Agarwal [35] proposed an approach to remedy 3-limitations: 1) machine translation, 2) online machine translation, and 3) the ability to identify different types of plagiarism by using machine learning and crowd-sourcing techniques. The results reported that LSI

works well in the field of multilingual retrieval. However, it needs some customization to be used in multilingual plagiarism.

Comparable corpora-based models. This model refers to documents that are translations of each other and whose words or sentences have been mapped manually or heuristically to their respective translations, e.g., Explicit Semantic Analysis (CL-ESA).

In general, it was found that out of the 81 papers that applied a CLPD technique only, 27 (33%) studies applied T+MA (MT-based models); to Dictionary-based models: 9 (11%) studies presented CL-CTS, and 5 (6%) studies applied CL-VSM; 7 (9%) studies used CL-ASA and 1 (1%) studies used CL-LSI both in Parallel corpora-based models; 5 (6%) CL-KGA and 1 (1%) study CL-ESA in Comparable corpora-based models; 3 (4%) studies utilized CL-CNG (Syntax-based models). At the same time, we also identified other models/techniques presented in selected studies, e.g., 17 (20%) adopted Word Embeddings Similarity (CL-WE) and Latent Semantic Analysis (CL-LSA) 2 (1%) studies in Semantic-based models; and 5 (6%) studies used CL-FUZZY (Fuzzy-based models).

Furthermore, we identified 12 studies that compare their approach against others in order to evaluate their performance with each other, e.g., CL-CNG vs. CL- VSM (2 studies) [36], [37]; CL-CTS vs. CL-CNG and CL-ASA (1 study) [38]; T+MA vs. CL- CNG and CL-ASA (3 studies) [4], [39], [40]; CL-CNG vs. CL-CTS vs. CL-ASA vs. CL-ESA vs. T+MA (1 study) [41]; CL-CNG vs. CL-ASA vs. T+MA (1 study) [42]; CL-KGA vs. CL-FUZZY (1 study) [43]; and combined to obtain a better performance: CL-CNG, CL-CTS, CL-WE, and T+MA (1 study) [44]; CL-WE and T+MA (1 study) [45]; Bilingual dictionary and T+MA (1 study) [46]; and 11 studies were related to building "Corpus" for CLPD methods.

Consequently, in terms of popularity of usage, T+MA (MT-based models) and CL-WE (Semantic-based models) were the most used CLPD models, while CL-LSI and CL-ESA models were the least used.

2) RQ2. How are the proposed techniques evaluated?

To address RQ2, we have determined the assessment criteria to understand how each technique was validated. We have assessed the validation of each technique with four artifacts: thesaurus, corpus, dataset, and others (documents, thesis, academic papers). For this study, we identified that "corpus" was the artifact most utilized to evaluate the proposed techniques for the cross-language plagiarism detection process (TABLE IV).

Thesaurus. In this study, few papers utilized thesaurus to verify CLPD methods. Some studies [19], [32], [47] discussed thesaurus in a broad sense. They outlined Eurovoc Conceptual Thesaurus to conduct a high similarity search. The authors denoted that many complex word structures are contained in a Conceptual Thesaurus, and every commonly used idea from the field is exhaustively covered. They also provided a concept that Eurovoc1 results from EU Parliamentary deliberations. They supported Eurovoc as it is a living resource containing nearly 6,797 multilingual concepts in twenty-two dialects labeled using concept IDs in the European Parliament. The authors preferred the model as a means of verification because it can apply across corpora, is stable and consistent, and exudes competitive outcomes. However, they argued that the tool is quite "generic".

Corpus. Out of the total number of analyzed papers, 42 (52%) agreed that Corpus can be used to evaluate the CLPD.

TABLE IV
ARTIFACTS

Artifact	Frequency	Example
Corpus	42	ECLaPA, German and English Wikipedia collections JRC-Acquis, PAN-09, PAN-10, PAN-PC-2011 corpus PAN-PC-12 text alignment corpus (German-English) PAN-PC-12 corpus, INTERSECT and OPUS corpus CL!TR 2011 corpus, UN multilingual corpora English-Persian Mizan parallel corpus, Europarl corpus, BabelNet English Chinese bilingual parallel corpus Google AJAX Search API (Google search engine) as corpus CLC_EVC- English-Vietnamese bilingual corpus SemCor, CLEU Corpus, TREU Corpus, CQADupStack corpus English-Persian bilingual plagiarism detection corpus A bilingual scientific publication corpus
Dataset	22	Human-rated benchmark dataset, CLiTR-Dataset SemEval-2017 STS, ASKUbuntu & Stack Overflow QATweets in the Swiss-German language, PAN2014 datasets
Other	22	Documents and their summaries Czech National Archive of Graduate Theses Collection contains a plain text files Documents on the web repository Academic documents & Scientific papers SPARQL queries over DBpedia, Encripted data High-volume multilingual text data Pairs of Arabic-English parallel sentences German-English and Spanish-English language partitions Post data annotated for errors in three language pairs
Thesaurus	4	Eurovoc; Eurovoc1 EuroWordNet

In other words, it was the best evaluation way that was registered in this study. This was the best evaluation that was

identified in this study. Loginova et al. [48] argued in their study that 'an accessible natural language interface' also includes multilingual question answering (MLQA). The authors also argued that current solutions show significant performance deficits compared to a single-language system, and examined a few different machine learning models, and they discovered that deep learning approaches improved MLQA performance considerably.

They compared 'performance of a deep learning model before and after' non-factored questions and answers were translated using corpus translation. Blloshmi et al. [49] complimented the corpus by using annotation projection; the study of Issa et al. [50] processed parallel phrases in the Europarl corpus to generate cross-lingual silver AMR annotations. In addition, they used a parallel corpus to relate English sentences to target language sentences by constructing an AMR graph for each English. Using an existing AMR parser, English sentences from the 'parallel corpus' were parsed and assigned their results to a 'PARSENTS-SILVERAMR' method. This study found that Asghari et al. [51] created an 'English-Persian bilingual plagiarism detection corpus' (also known as HAMTA-CL) to support the evaluation of' cross-language plagiarism detection approaches'. It included seven types of obfuscation to help in the evaluation of cross-language plagiarism detection approaches Simple Translation, Artificial, Paraphrasing, Summarization, Circular translation, Split, and Merge obfuscation. One of the limitations of evaluating and comparing the effectiveness of systems used for cross-language plagiarism detection is becoming increasingly evident [52]. As a result, the use of corpus was significantly increased in this work. Chang et al. [53] reported that prior studies used enormous bilingual resources as references, including 'parallel corpus', 'comparable corpus', 'and even commercial machine translation'. Torrejón & Ramos considered these verification techniques to be inconvenient, time-consuming, and impractical [20], [54]. In this regard, they developed CL-WMD, which, surprisingly, still relied on the corpus of scholarly publications. This means that corpus (dictionaries) will continue to have verification influence even if a new CLPD is developed. Nonetheless, other verification methods garnered 13% of the total papers analyzed. It was less significant as the percentage came due to various small CLPD verification methods other than a single one across the studies [55].

Datasets. 22 (27%) selected studies cited "dataset" as a CLPD artifact. Ehsan & Shakery suggested a topic-based segmentation approach in their paper 'Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information' to transform the suspicious text into a series of related passages, which are essentially datasets [56]. Gupta and Singhal [29] also used datasets to determine the most likely English source document for a given Hindi dubious text. Pereira et al. [57] also employed the data mining technology 'Weka data' (datasets) to detect plagiarism in the PAN'10 competition. They argued that CLPD consists essentially of five phases: 'linguistic normalization, document retrieval, classifier training, plagiarism analysis, and post-processing.' Other researchers endorsed the conclusions about datasets [6], [58]–[60].

3) RQ3. What is the available support for the identified techniques?

This study also assesses the available support for executing a cross-language plagiarism detection technique. This review identified that the cross-language plagiarism detection process is automated, as we expected. However, few selected studies describe its tool. The environment selected for its deployment was: Web (8 studies), Desktop (2 studies), and Grid (2 studies). All these tools were developed for academic purpose.

Web. 8 studies reported using the web to conduct CLPD detection. Rücklé et al. [61] found that "the standard approach to cross-lingual information retrieval, which automatically translates the query to the target language and continues with a monolingual retrieval model, typically falls short in cQA due to translation errors". Hence, Google Translate became their tool because it was free of the various faults that other technical cQA solutions produced. This means that using the internet as a support for running CLPD was sufficient here. While ArbEngVec (in place of the web) was utilized by [62] to conduct Information Retrieval (IR) and Information Extraction (IE), the group used it in place of the web to implement multiple Arabic-English cross-lingual word embedding models. Thereby, CLPD execution is mapped to ArbEngVec. According to this source, the researchers who carried out the study, Shojaie and Safi-Esfahani [63] found that ParaMaker, capable of generating precise paraphrases of any sentence, is comparable to human behavior transmitting those to a search engine to find plagiarism patterns, was employed. The Paramaker tool outperformed the other tools by 34% when it came to finding similarity indexes. The study of Bakhteev et al. [64] also suggested that web tools are utilized for the execution of CLPD instructions, in particular, the CrossLang system for English-Russian language pairs, which uses plagiarism detection. CrossLang should be added to the list of other web tools that can be used to detect CLPD.

Desktop/grid application. Guan et al. [65] provided the results of their study, which finds that the appearance of searchable encryption technology brings new focus to the challenge of discovering encrypted data with secure search. They discovered an entirely new approach to searching over encrypted cloud data by devising an individual solution to the cross-lingual search problem [66]. A cross-lingual multi-keyword rank search algorithm called CLRSE (Cross-Lingual Multi-Keyword Rank Search) was created based on the Open Multilingual Wordnet and helps break down language barriers while increasing speed and functionality of search. Because of this, Wordnet serves both as a desktop program and a cloud system to assist CLPD work [54], [67]. Developed a double purpose to help with CLPD implementation on both the desktop and cloud platforms. Because of the higher computational complexity, the winnowing technique developed by the Electrical Engineering Department, Universitas Indonesia, for cross-language plagiarism detection was unusable in the actual world. As a result, they worked to see if similar systems on a lab-scale multicore-based private cloud platform called OpenStack might be parallelized. If compared the time it took to complete the serial computation (execution time) to the time it took to speed up parallelization to 3.52 times, they completed the parallel computation faster (in the event). Since CLPD's tools

have been found to support functions both on the computer and in the clouds, the Open stack tool has been selected to help CLPD carry out this task [23], [68].

B. Discussion

This section presents the main findings of this systematic literature review. This SLR focuses on how cross-language plagiarism detection techniques are employed in the Natural Language context. To maximize coverage of potentially relevant studies retrieval and ensure that these SLR results cover all studies that present cross-language plagiarism detection techniques, we used the terms "cross-language", "plagiarism" and "detection" as part of the search terms for this SLR. Cross-Language Plagiarism Detection techniques are carried out depending on the kind of artifact (e.g., thesaurus, corpus, dataset) and/or aspects in the source language of the suspicious documents. However, the aspects are given by the grammatical particularity (syntax, scripts, word order, use of verb tenses, etc.) of the language itself, e.g., [8], [33]; and researchers have needed to deal with these characteristics to obtain better performances in their plagiarism detection systems through many approaches, but no single technique is suitable for all circumstances and types of artifacts. It depends on the evaluation purpose and the kind of artifact that is evaluated.

RQ1. Techniques are employed for Cross-Language Plagiarism detection (CLPD). This question focused much of its attention on finding out the techniques that are applied in CLPD. Some of the techniques that arose from the study include Translation-based and monolingual analysis (T+MA) [21], [24], Dictionary and thesaurus-based approaches [18], [28], [29], Parallel corpora-based models [33], [35], Comparable corpora-based models, Semantic-based models, etc. Among these models, Semantic-based models (CL-WE) took the lead in usage across the studied papers. This raises questions about why Semantic-based Models may be chosen over others. Furthermore, combining methods, e.g., CL-WE and T+MA [6], could provide better yields. While the noisiness and informal nature of the social media genre present additional challenges to cross-lingual embedding methods, they also provide opportunities because of the abundance of code-switching and the existence of a shared vocabulary of emoji and named entities [69]. Levy and Wang [70] proposed in favor of semantic-based techniques for CLPD because of pandemics like COVID-19, which was more like a small-scale pandemic. They assert that the widespread use of COVID-19 has been a large and damaging problem in society by 2020. New epidemics had occurred across the world, and they followed previously impacted regions.

Many illness detection algorithms lack a social media data pool from which useful modeling and prediction information might be obtained. While this applied, the knowledge that was gathered here was important for the two to test whether that knowledge could be used to mimic an outbreak in another country. To align different language populations for epidemiological purposes, they recommended the use of cross-lingual transfer learning. By training on Italy's early COVID-19 epidemic on Twitter and transfer to other countries. Levy and Wang [70] employed both macro and micro text elements with a correlation of as much as 0.85. The studies indicated promising results for cross-country forecasts. In addition to previous studies, Nguyen and Dien [43] utilized multilingual word embedding, POS vectorization, and a POS tagging label redesign to support Semantic-based CLPD models. After utilizing the semantic models, the models' accuracy was increased from 86.86% to 89.61%. Using semantic models, Poerner and Schütze [71] analyzed the problem of recognizing duplicate questions (DQD). An international team of researchers developed a Semantic-based Multilingual BERT model that distinguished informative and actionable tweets on Twitter to improve disaster management. The constant definition refers to the idea that all of the publications below (namely, papers in this collection and elsewhere) state that semantic-based CLPD identification models, such as CL-WE, are the most effective in solving plagiarism incidents in many fields and platforms. Alzahrani et al. [3] found that most plagiarism detection algorithms are ineffective since they concern themselves with plagiarism detection only based on copying and pasting text with/without small changes to the language and grammar [72]. Most algorithms fail to identify plagiarism because they paraphrase, summarize, and keep the same idea while copying others' ideas and contributions. When one reads a pirated text published uniquely, they will see why several existing plagiarism detection methods do not consider the crossover [4]. For instance, they discovered several new things when Cedeño [19] evaluated three cross-language similarity analysis techniques. First, despite Barrón-best Cedeno's efforts, CL-ESA and MLPlag were left out of the comparison. Their research was focused on detecting sentence-level plagiarism, and MLPlag was created to assess complete papers. On the other hand, Cedeño [19] stated that CL-ASA outperforms when used in languages with disparate alphabets or syntax, as concluded by Cedeño [2]. Language pairings that end in "eu" and "es-eu" are precisely the ones mentioned above. Additional studies indicated that Basque Wikipedia pages do not meet the standards for a similar corpus, as reported by [19]. Still, the results from the study conducted by [19], which agree with this study's findings, stated that although the strategy of using T+MA is simple, it performs well for en-eu and es-eu. The T+MA approach also proves superior since CL-ASA has less of an impact on the lack of resources. This could be because it takes into consideration both the e[n|s] model (e[n|s]-eu) and the eu e[n|s] model (eu e[n|s]).

Another advantage of the language model is that it minimizes incorrect translations while including additional information about the syntax. It is exactly the opposite; CL-ASA ignores syntactic links between the texts entirely. To achieve a better outcome, one must invest more in computing resources. It is just necessary to perform string comparisons for CL-CNG. The key requirement for CL-ASA is that translation probabilities must be present in aligned corpora, but once this has been done, cross-language similarity may be calculated quickly. However, according to Cedeño [19], T+MA can be highly costly for big collections since the preceding translation of all the texts is required. Other authors also gave compelling reasons why T+MA could be the best CLPD technique. For instance, Rosso [72] discovered that the resulting dictionary was inadequate in a study where he brought aboard the disadvantages of doing CLPD on a tiny

language like Amazigh. T+MA emerged as the most effective option. After interpreting all the resources into the local dialect, no amount of computing effort would solve the challenge of cross-language plagiarism detection without being monolingual. Additionally, even though Alzahrani [25] and Safi-Esfahani et al. [26] conducted CLPD with a semantic method, they used the basics of monolingual to execute the multilingual and cross-lingual plagiarism detection. It means that T+MA appears as a baseline for conducting any plagiarism detection. The study done by Ferrero et al. [41] revealed the efficacy and versatility of the T+MA approach. They stated that CL-ESA appears to produce better outcomes on homologous corpora, such as Wikipedia. Ferrero and his colleagues also asserted that CL-ASA fares better on parallel corpora like JRC, Europarl, and APR. CL-C3G is the most efficient technique if the corpus contains named entities. They also discovered that CL-CTS and T+MA are super advantageous and adaptable. As a final note, the researchers found that CL-ESA was not very effective; it is the most time-consuming technique, and it is highly reliant upon the corpus employed. Other authors that supported the approach also included [6], [54], [55], [73], among others.

Finally, we have found no defined rules or guidelines for applying a certain cross-language plagiarism detection technique. Some techniques base their analysis on information retrieval, either by scraping web pages or downloading a data dump from the internet. However, this information might be wrong, e.g., Wikipedia [74]. Moreover, few authors present their process using a tool, yet it is not available anymore. Therefore, the lack of guidelines and tools may affect the techniques' ability to ensure reproducibility and thus the quality of results.

RQ2: CLPD Evaluation artifacts. This question was directed towards identifying some of the methods of evaluating CLPD identification models. From the study, the following artifacts surfaced: thesaurus and corpus. Many researchers noted that many complex word structures are contained in a conceptual thesaurus, and every commonly used idea in the field is exhaustively covered, which makes it a difficult tool to navigate. The study found that the most used artifact for evaluating the models was the Corpus (dictionaries). Natural Language Processing Techniques and Fuzzy Semantic Similarity for Automatic External Plagiarism Detection were examined by Gupta et al. [75]. PAN-2012, PAN2010, and PAN-09 corpora were employed in the analysis. The detection results were promising despite the procedure being computationally intensive. Asghari et al. [51] carried out a cross-language plagiarism detection study and developed an English-Persian bilingual plagiarism detection corpus that included seven obfuscation categories. These studies by Chang et al. [53] and Zubarev and Sochenkov [76], and other researchers who contributed to the studies' publications all relied on corpora as part of their methods for ascertaining the CLPD. According to Asghari et al. [51], the main use of the corpus is due to the wide variety of languages it covers and which prevents the many common case errors that translate across different languages using Google Translate. The argument that corpus is the best CLPD evaluation method also found support from some studies [55], [58], [73], [77]. Artifact validation may vary, but they are all in the natural language context, and it is a topic that is still growing. However, cross-language plagiarism detection techniques seem to be used in some way independent of the context. These techniques might also be applied in the Software Engineering domain within online communities, e.g., Stack Overflow (SO). SO is a question-and-answer website to help developers in English, Portuguese, Spanish, Russian and Japanese, which combines natural language and source code [78].

Consequently, it will be interesting to know whether some techniques may be applied in SO in order to avoid cross-language post duplicates among all SO websites.

RQ3: Available CLPD support. This question assessed the available support for executing cross-language plagiarism detection techniques. It was found that web-based tools, desktop, and grid/cloud tools provided significant support for CLPD execution. Nonetheless, web-based support tools were identified as the most relevant support tools for the CLPD methods. It goes without saying that we are living in a digital era. It is an era engraved with numerous online searches, ad hoc library visits, and remote research and working. Never forgetting the ongoing world pandemic COVID-19, people are forced to avoid movement, work from home, and even avoid visiting computer hubs. It remains an option to venture into online data mining and research. Thus, in dealing with plagiarism detection, web-based verification tools could suffice in this world. The data collected in this study also agreed with the above statement that web tools were the best support available for CLPD execution. Gipp et al., used CitePlag [79], a prototype of a PDS that merges citation trend analysis with conventional personality testing methods from their prior study to depict the most extended instance of cross-language plagiarism in Guttenberg's thesis. To provide users with more interactive and straightforward document comparison, they can customize the highlights of citation-based and character-based similarity information to make it more visually appealing. They discovered that the online application was critical to the success of CLPD. The findings were supported by some studies [24], [57], [80]. Finally, considerable efforts have been applied to detecting cross-language plagiarism, and although some authors [2]–[6], [8]–[10], [81] have proposed or identified techniques capable of detecting.

C. Threats to Validity

We consider internal and external threats to validity [11].
External validity. We chose Google Scholar as the source of publication and selection bias, where papers about cross-language plagiarism detection techniques commonly appear. We compared the retrieved documents against a small sample that was previously identified as relevant papers to study. However, we did not consider gray literature (e.g., technical reports, Ph.D. thesis) or unpublished results.

Internal validity. In the search string, we tried to collect all the strings that are representative of the research question. We redefined our search string to achieve the maximum of papers related to the systematic review. Besides, we have taken into consideration synonyms and have included lexical words in our words. Finally, we attempted to alleviate the threat of inaccuracy in data extraction and misclassification by conducting the classifications of the papers with three reviewers and solving the discrepancies by consensus.

IV. Conclusion

This study presents a systematic literature review to determine what cross language plagiarism detection techniques have been employed in the Natural Language context. We started with 130K potentially relevant studies, and after applying our inclusion/exclusion criteria and snowballing, we concluded our SLR with 107 documents. We provided an overview of different techniques to tackle cross-language plagiarism detection. Our results show that most of the selected studies follow a Machine Translation + Monolingual Analysis (T+MA). It means they first translated the suspicious and original documents into a common language (e.g., English) and then applied any monolingual techniques, e.g., fingerprints.

Many of the selected studies used "corpus" as the artifact in the evaluation phase. And we also observed there is no specific supporting tool; a few papers present some minimum details about its generic tool. Finally, we identified the implementation of any of these cross-language plagiarism detection techniques in a Software Engineering context as future work. Online communities like Stack Overflow present the problem of cross-language posts among SO websites.

Acknowledgment

This research was supported by the SENESCYT-Ecuador (scholar- ship program 2013-2).

References

[1] IEEE, "A Plagiarism FAQ," 2015. [Online]. Available: http://www.ieee.org/publications_standards/publications/rights/plagiarism_FAQ.html. [Accessed: 11-May-2018].

[2] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection," in *Coling 2010: Posters*, 2010, pp. 997–1005.

[3] S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 2. pp. 133–149, 2012.

[4] A. Barrón-Cedeño, P. Gupta, and P. Rosso, "Methods for cross-language plagiarism detection," *Knowledge-Based Syst.*, vol. 50, pp. 211–217, 2013.

[5] M. Franco-Salvador, P. Rosso, and M. Montes-y-Gómez, "A systematic study of knowledge graph analysis for cross-language plagiarism detection," *Inf. Process. Manag.*, vol. 52, no. 4, pp. 550–570, 2016.

[6] J. Ferrero, L. Besacier, D. Schwab, and F. Agnès, "Deep Investigation of Cross-Language Plagiarism Detection Methods," in *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, 2017, pp. 6–15.

[7] A. E. Tlitova, A. S. Toschev, M. Talanov, and V. Kurnosov, "Meta-Analysis of Cross-Language Plagiarism and Self-Plagiarism Detection Methods for Russian-English Language Pair.," *Front. Comput. Sci.*, vol. 2, p. 523053, 2020.

[8] A. Kumar and S. Das, "An evolutionary survey from Monolingual Text Reuse to Cross Lingual Text Reuse in context to English-Hindi," *Int. J. Sci. Eng. Res.*, vol. 6, no. 2, pp. 996–1003, 2015.

[9] S. Shimpikar and S. Govilkar, "A Survey of Text Summarization Techniques for Indian Regional Languages," *Int. J. Comput. Appl.*, vol. 165, no. 11, pp. 29–33, 2017.

[10] P. Rosso, "Author profiling and Plagiarism detection," in *Communications in Computer and Information Science*, 2015, vol. 505, pp. 229–250.

[11] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, 2014, pp. 1–10.

[12] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS Q.*, pp. xiii--xxiii, 2002.

[13] Keele University, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[14] A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E. D. López-Cózar, "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations," *Scientometrics*, vol. 126, no. 1. Springer, pp. 907–908, 2021.

[15] D. Landman, A. Serebrenik, and J. J. Vinju, "Challenges for static analysis of java reflection-literature review and empirical study," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 2017, pp. 507–518.

[16] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[17] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[18] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Coling 2010: Posters*, 2010, pp. 36–44.

[19] A. Barrón Cedeño, "On the Mono- and cross-language detection of text re-use and plagiarism," *Proces. Leng. Nat.*, vol. 50, pp. 103–105, 2013.

[20] M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso, "Cross-language plagiarism detection," *Language Resources and Evaluation*, vol. 45, no. 1. pp. 45–62, 2011.

[21] J. Kasprzak and M. Brandejs, "Improving the reliability of the plagiarism detection system: Lab report for PAN at CLEF 2010," in *CEUR Workshop Proceedings*, 2010, vol. 1176, pp. 359–366.

[22] C. K. Kent and N. Salim, "Web based cross language plagiarism detection," in *2010 Second International Conference on Computational Intelligence, Modelling and Simulation*, 2010, pp. 199–204.

[23] M. Pataki, "A new approach for searching translated plagiarism," 2012.

[24] C. K. Kent and N. Salim, "Web based cross language semantic plagiarism detection," in *Proceedings - IEEE 9th International Conference on Dependable, Autonomic and Secure Computing, DASC 2011*, 2011, pp. 1096–1102.

[25] S. Alzahrani, "Cross-Language Semantic Similarity of Arabic-English Short Phrases and Sentences.," *J. Comput. Sci.*, vol. 12, no. 1, pp. 1–18, 2016.

[26] F. Safi-Esfahani, S. Rakian, and M. H. Nadimi-Shahraki, "English-Persian Plagiarism Detection based on a Semantic Approach," *J. AI Data Min.*, vol. 5, no. 2, pp. 275–284, 2017.

[27] R. Kothwal and V. Varma, "Cross lingual text reuse detection based on keyphrase extraction and similarity measures," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7536 LNCS, pp. 71–78.

[28] P. Gupta, K. Singhal, P. Majumder, and P. Rosso, "Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism," *ICON*, 2011.

[29] P. Gupta and K. Singhal, "Mapping Hindi-English text re-use document pairs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7536 LNCS, pp. 79–85.

[30] T. Brychcín, "Linear transformations for cross-lingual semantic textual similarity," *Knowledge-Based Syst.*, vol. 187, p. 104819, 2020.

[31] Z. Ceska, M. Toman, and K. Jezek, "Multilingual plagiarism detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5253 LNAI, pp. 83–92.

[32] P. Gupta, A. Barrón-Cedeño, and P. Rosso, "Cross-language high similarity search using a conceptual thesaurus," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7488 LNCS, pp. 67–75.

[33] D. Pinto, J. Civera, A. Barrón-Cedeno, A. Juan, and P. Rosso, "A statistical approach to crosslingual natural language tasks," *J. Algorithms*, vol. 64, no. 1, pp. 51–60, 2009.

[34] S. Yahyaei, M. Bonzanini, and T. Roelleke, "Cross-lingual text fragment alignment using divergence from randomness," in *International Symposium on String Processing and Information Retrieval*, 2011, pp. 14–25.

[35] M. Mostafa and L. Agarwal, "Multilingual Plagiarism Detection," 2014.

[36] S. Alzahrani, N. Salim, A. A.-I. T. on, and undefined 2011, "Understanding plagiarism linguistic patterns, textual features and detection methods," *researchgate.net*.

[37] N. Ehsan, F. Tompa, … A. S. the 2016 A. S. on, and undefined 2016, "Using a dictionary and n-gram alignment to improve fine-grained cross-language plagiarism detection," *dl.acm.org*.

[38] P. Gupta, A. Barrón-Cedeno, and P. Rosso, "Cross-language high similarity search using a conceptual thesaurus," *Conf. Cross-Language*, 2012.

[39] A. Barrón-Cedeno, P. Rosso, and E. Agirre, "Plagiarism detection across distant language pairs," *Proc. 23rd*, 2010.

[40] J. Ferrero, L. Besacier, D. Schwab, and F. Agnes, "Deep Investigation of Cross-Language Plagiarism Detection Methods," in *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, 2017, pp. 6–15.

[41] J. Ferrero, F. Agnès, L. Besacier, and D. Schwab, "A multilingual, multi-style and multi-granularity dataset for cross-language textual similarity detection," in *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2016, pp. 4162–4169.

[42] M. Potthast, A. Barrón-Cedeño, and B. Stein, "Cross-language plagiarism detection," *Lang. Resour.*, 2011.

[43] L. T. Nguyen and D. Dien, "Vietnamese- English Cross-Lingual Paraphrase Identification Using Siamese Recurrent Architectures," in *Proceedings - 2019 19th International Symposium on Communications and Information Technologies, ISCIT 2019*, 2019, pp. 70–75.

[44] J. Ferrero, F. Agnes, L. Besacier, and D. Schwab, "CompiLIG at SemEval-2017 Task 1: Cross-language plagiarism detection methods for semantic textual similarity," *arXiv Prepr. arXiv1704.01346*, 2017.

[45] E. M. B. Nagoudi, J. Ferrero, D. Schwab, and H. Cherroun, "Word embedding-based approaches for measuring semantic similarity of arabic-english sentences," in *International Conference on Arabic Language Processing*, 2017, pp. 19–33.

[46] H. Ezzikouri, M. Erritali, and M. Oukessou, "Plagiarism Detection in Across Less Related Languages (English-Arabic): A Comparative Study," in *Smart Data and Computational Intelligence*, 2019, pp. 207–213.

[47] C. Vania and M. Adriani, "Automatic external plagiarism detection using passage similarities," in *CEUR Workshop Proceedings*, 2010, vol. 1176.

[48] E. Loginova, S. Varanasi, and G. Neumann, "Towards End-to-End Multilingual Question Answering," *Inf. Syst. Front.*, vol. 23, no. 1, pp. 227–241, 2021.

[49] R. Blloshmi, R. Tripodi, and R. Navigli, "XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020, pp. 2487–2500.

[50] F. Issa, M. Damonte, S. B. Cohen, X. Yan, and Y. Chang, "Abstract meaning representation for paraphrase detection," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018, vol. 1, pp. 442–452.

[51] H. Asghari, O. Fatemi, S. Mohtaj, H. Faili, and P. Rosso, "On the use of word embedding for cross language plagiarism detection," *Intell. Data Anal.*, vol. 23, no. 3, pp. 661–680, 2019.

[52] A. Micsik, P. Pallinger, and D. Siklósi, "Scaling a Plagiarism search service on the BonFIRE testbed," in *Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom*, 2013, vol. 2, pp. 57–62.

[53] C. Chang, C.-H. Chang, and S.-Y. Hwang, "Employing word mover's distance for cross-lingual plagiarized text detection," *Proc. Assoc. Inf. Sci. Technol.*, vol. 57, no. 1, p. e229, 2020.

[54] D. A. R. Torrejón, J. Manuel, and M. Ramos, "Detailed Comparison Module In CoReMo 1 . 9 Plagiarism Detector Notebook for PAN at CLEF 2012," *CLEF (Online Work. Notes/Labs/Workshop)*, pp. 1–8, 2012.

[55] A. P. Zakiy Firdaus Alfikr, "The Construction of Indonesian-English Cross Language Plagiarism Detection System Using Fingerprinting Technique," *J. Comput. Sci. Inf.*, vol. 5, no. 1, pp. 16–23, 2012.

[56] N. Ehsan and A. Shakery, "Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information," *Inf. Process. Manag.*, vol. 52, no. 6, pp. 1004–1017, 2016.

[57] R. C. Pereira, V. P. Moreira, and R. Galante, "UFRGS @ PAN2010 : Detecting External Plagiarism," in *Lab Report for PAN at CLEF 2010*, 2010.

[58] L. Gang, Z. Quan, and L. Guang, "Cross-language plagiarism detection based on WordNet," in *ACM International Conference Proceeding Series*, 2018, vol. Part F1376, pp. 163–168.

[59] K. Mustofa and Y. A. Sir, "Early-Detection system for cross-language (translated) plagiarism," in *Information and Communication Technology-EurAsia Conference*, 2013, pp. 21–30.

[60] L. T. Nguyen and D. Dien, "English-Vietnamese cross-language paraphrase identification method," in *ACM International Conference Proceeding Series*, 2017, vol. 2017-Decem, pp. 42–49.

[61] A. Rücklé, N. S. Moosavi, and I. Gurevych, "Neural duplicate question detection without labeled training data," *arXiv Prepr. arXiv1911.05594*, 2019.

[62] R. Lachraf, Y. Ayachi, A. Abdelali, D. Schwab, and others, "ArbEngVec: Arabic-English cross-lingual word embedding model," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 2019, pp. 40–48.

[63] A. Shojaei and F. Safi-Esfahani, "External Plagiarism Detection based on Human Behaviors in Producing Paraphrases of Sentences in English and Persian Languages," *J. AI Data Min.*, vol. 7, no. 3, pp. 451–466, 2019.

[64] O. Bakhteev, A. Ogaltsov, A. Khazov, K. Safin, and R. Kuznetsova, "CrossLang: the system of cross-lingual plagiarism detection," *Work. Doc. Intell. NeurIPS 2019*, no. 18, pp. 1–5, 2019.

[65] Z. Guan *et al.*, "Cross-lingual multi-keyword rank search with semantic extension over encrypted data," *Inf. Sci. (Ny).*, vol. 514, pp. 523–540, 2020.

[66] M. Franco-Salvador, P. Rosso, and R. Navigli, "A knowledge-based representation for cross-language document retrieval and categorization," in *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, 2014, pp. 414–423.

[67] A. A. Putri Ratna, F. Astha Ekadiyanto, I. Ibrahim, D. Husna, and F. Rahimullah, "Investigating Parallelization of Cross-language Plagiarism Detection System Using the Winnowing Algorithm in Cloud Based Implementation," in *2019 IEEE 10th International Conference on Awareness Science and Technology, iCAST 2019 - Proceedings*, 2019, pp. 1–7.

[68] M. Pataki and A. C. Marosi, "Searching for Translated Plagiarism with the Help of Desktop Grids," *J. Grid Comput.*, vol. 11, no. 1, pp. 149–166, 2013.

[69] J. Camacho-Collados, Y. Doval, E. Martínez-Cámara, L. Espinosa-Anke, F. Barbieri, and S. Schockaert, "Learning cross-lingualword embeddings from Twitter via distant supervision," in *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, 2020, vol. 14, pp. 72–82.

[70] S. Levy and W. Y. Wang, "Cross-lingual Transfer Learning for COVID-19 Outbreak Alignment," *arXiv Prepr. arXiv2006.03202*, 2020.

[71] N. Poerner and H. Schütze, "Multi-view domain adapted sentence embeddings for low-resource unsupervised duplicate question detection," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2020, pp. 1630–1641.

[72] P. Rosso, "On the risk of cross-language plagiarism for less resourced languages such as Amazigh," *users.dsic.upv.es*, vol. 5, pp. 53–70, 2008.

[73] I. Muneer, M. Sharjeel, M. Iqbal, R. M. A. Nawab, and P. Rayson, "CLEU - A Cross-language english-urdu corpus and benchmark for text reuse experiments," *J. Assoc. Inf. Sci. Technol.*, vol. 70, no. 7, pp. 729–741, 2019.

[74] M. Potthast, B. Stein, and M. Anderka, "A Wikipedia-based multilingual retrieval model," in *European conference on information retrieval*, 2008, pp. 522–530.

[75] D. Gupta, K. Vani, and C. K. Singh, "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 2014*, 2014, pp. 2694–2699.

[76] D. V Zubarev and I. V Sochenkov, "Cross-language text alignment for plagiarism detection based on contextual and context-free models," in *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 2019, vol. 2019-May, no. 18, pp. 809–820.

[77] N. Ehsan, A. Shakery, and F. W. Tompa, "Cross-lingual text alignment for fine-grained plagiarism detection," *J. Inf. Sci.*, vol. 45, no. 4, pp. 443–459, 2019.

[78] M. Botto-Tobar, W. Torres, A. Lozano, M. G. J. van den Brand, B. Vasilescu, and A. Serebrenik, "Is stack overflow in portuguese attractive for brazilian users?," in *Proceedings of the 13th*

International Conference on Global Software Engineering, 2018, pp. 21–29.

[79] B. Gipp, N. Meuschke, C. Breitinger, J. Pitman, and A. Nürnberger, "Web-based demonstration of semantic similarity detection using citation pattern visualization for a cross language plagiarism case," in *ICEIS 2014 - Proceedings of the 16th International Conference on Enterprise Information Systems*, 2014, vol. 2, pp. 677–683.

[80] S. Alzahrani, N. Salim, C. K. Kent, M. S. Binwahlan, and L. Suanmali, "The development of cross-language plagiarism detection tool utilising fuzzy swarm-based summarisation," in *Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, ISDA'10*, 2010, pp. 86–90.

[81] H. Ezzikouri, M. Erritali, and M. Oukessou, "Plagiarism Detection in Across Less Related Languages (English-Arabic): A Comparative Study," in *International Conference on Advanced Information Technology, Services and Systems*, 2018, pp. 207–213.

[82] S. Parida, E. Villatoro-Tello, S. Kumar, P. Motlicek, and Q. Zhan, "Idiap Submission to Swiss-German Language Detection Shared Task.," in *SwissText/KONVENS*, 2020.

[83] M. Roostaee, M. H. Sadreddini, and S. M. Fakhrahmad, "An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes," *Inf. Process. \& Manag.*, vol. 57, no. 2, p. 102150, 2020.

[84] A. Shojaie and F. Safi-Esfahani, "External Plagiarism Detection based on Human Behaviors in Producing Paraphrases of Sentences in English and Persian Languages," *J. AI Data Min.*, vol. 7, no. 3, pp. 451–466, 2019.

[85] F. Ture, T. Elsayed, and J. Lin, "No free lunch: brute force vs. locality-sensitive hashing for cross-lingual pairwise similarity," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 943–952.

[86] R. Pereira, V. Moreira, and R. Galante, "A new approach for cross-language plagiarism analysis," *Conf. Cross-Language*, 2010.

[87] B. Pouliquen, R. Steinberger, and C. Ignat, "Automatic identification of document translations in large multilingual document collections," *arXiv Prepr. cs/0609060*, 2006.

[88] Z. Ceska, M. Toman, and K. Jezek, "Multilingual plagiarism detection," *Int. Conf. Artif.*, 2008.

[89] V. Thompson, "Detecting cross-lingual plagiarism using simulated word embeddings," *arXiv Prepr. arXiv1712.10190*, 2017.

[90] J. Ray Chowdhury, C. Caragea, and D. Caragea, "Cross-lingual disaster-related multi-label tweet classification with manifold mixup," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2020.

[91] S. Alzahrani and H. Aljuaid, "Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: A study on Arabic-English plagiarism cases," *J. King Saud Univ. Inf. Sci.*, 2020.

[92] C. Lo and M. Simard, "Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 206–215.

[93] A. K. Khakimova, M. M. Charnine, A. A. Klokov, and E. G. Sokolov, "Approaches to assessing the semantic similarity of texts in a multilingual space," 2020.

[94] N. Alotaibi and M. Joy, "Using Sentence Embedding for Cross-Language Plagiarism Detection," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2020, pp. 373–379.

[95] M. Ustaszewski, "Exploring Adequacy Errors in Neural Machine Translation with the Help of Cross-Language Aligned Word Embeddings," in *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, 2019, pp. 122–128.

[96] J. Ferrero, F. Agnes, L. Besacier, and D. Schwab, "Usingword embedding for cross-language plagiarism detection," *arXiv Prepr. arXiv1702.03082*, 2017.

[97] A. A. P. Ratna *et al.*, "Cross-language plagiarism detection system using latent semantic analysis and learning vector quantization," *Algorithms*, vol. 10, no. 2, p. 69, 2017.

[98] A. A. P. Ratna *et al.*, "Cross-Language Automatic Plagiarism Detector Using Latent Semantic Analysis and Self-Organizing Map," in *Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality*, 2018, pp. 83–87.

[99] S. Srivastava and S. Govilkar, "A Survey on Paraphrase Detection Techniques for Indian Regional Languages," *Int. J. Comput. Appl.*, vol. 163, no. 9, pp. 975–8887, 2017.

[100] M. S. Arefin, Y. Morimoto, and M. A. Sharif, "BAENPD: A Bilingual Plagiarism Detector.," *J. Comput.*, vol. 8, no. 5, pp. 1145–1156, 2013.

[101] M. Muhr and R. Kern, "External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system," in *2nd International Competition on Plagiarism Detection*, 2010.

[102] Y. Qin, "Cross-Lingual Similarity Discrimination with Translation Characteristics," *Int. J. Artif. Intell. \& Appl.*, vol. 4, no. 5, p. 39, 2013.

[103] A. Barrón-Cedeno, P. Rosso, D. Pinto, and A. Juan, "On Cross-lingual Plagiarism Analysis using a Statistical Model.," *PAN*, vol. 212, pp. 1–10, 2008.

[104] M. Franco-Salvador, P. Gupta, and P. Rosso, "Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing," in *PROMISE Winter School*, 2013, pp. 227–236.

[105] M. Franco-Salvador, P. Gupta, … P. R.-K.-B., and undefined 2016, "Cross-language plagiarism detection over continuous-space-and knowledge graph-based representations of language," *Elsevier*.

[106] N. Radoev, A. Zouq, and M. Gagnon, "Multilingual Question Answering using Lexico-Syntactic Patterns," *Resource*, vol. 65, pp. 86–88.

[107] M. Potthast, B. Stein, and M. Anderka, "A Wikipedia-based multilingual retrieval model," *Eur. Conf. Inf.*, 2008.

[108] H. Ezzikouri, M. Erritali, and M. Oukessou, "Fuzzy-semantic similarity for automatic multilingual plagiarism detection," *Int. J. Adv. Comput. Sci. Appl*, vol. 8, no. 9, pp. 86–90, 2017.

[109] H. Ezzikouri, M. Oukessou, M. Youness, and M. Erritali, "Fuzzy cross language plagiarism detection (Arabic-English) using WordNet in a big data environment," in *Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing*, 2018, pp. 22–27.

[110] D. Dinh and N. Le Thanh, "English--Vietnamese cross-language paraphrase identification using hybrid feature classes," *J. Heuristics*, pp. 1–17, 2019.

[111] A. Barrón-Cedeno, P. Rosso, S. Devi, and P. Clough, "Pan@ fire: Overview of the cross-language! ndian text re-use detection competition," *Inf. Access*, 2013.

[112] J. Kasprzak and M. Brandejs, "Improving the Reliability of the Plagiarism Detection System," in *Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF 2010), Uncovering Plagiarism, Authorship, and Social Software Misuse Worksop (PAN'10)*, 2010, pp. 359–366.

[113] R. Kothwal and V. Varma, "Cross lingual text reuse detection based on keyphrase extraction and similarity measures," *Multiling. Inf. Access South Asian*, 2013.

[114] D. A. R. Torrejón and J. M. M. Ramos, "Text alignment module in CoReMo 2.1 plagiarism detector," *Forner et al.[34]*, 2013.

[115] Z. Alaa, S. Tiun, and M. Abdulameer, "Cross-Language Plagiarism of Arabic-English Documents Using Linear Logistic Regression.," *J. Theor. \& Appl. Inf. Technol.*, vol. 83, no. 1, 2016.

[116] A. Aljohani and M. Mohd, "Arabic-English cross-language plagiarism detection using winnowing algorithm," *Inf. Technol. J.*, vol. 13, no. 14, p. 2349, 2014.

[117] M. Al-suhaiqi11, M. A. S. Hazaa22, and M. Albared33, "Arabic English Cross-Lingual Plagiarism Detection Based on Keyphrases Extraction, 2 Monolingual and Machine Learning Approach 3," 2018.

[118] M. Sharjeel, *Mono-and cross-lingual paraphrased text reuse and extrinsic plagiarism detection*. Lancaster University (United Kingdom), 2020.

[119] A. Rücklé, K. Swarnkar, and I. Gurevych, "Improved cross-lingual question retrieval for community question answering," in *The world wide web conference*, 2019, pp. 3179–3186.

[120] R. Jungnickel, A. Pomp, A. Kirmse, X. Li, V. Samsonov, and T. Meisen, "Evaluation and Comparison of Cross-lingual Text Processing Pipelines," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019, pp. 417–425.

[121] L. Gang, Z. Quan, and Y. Qianru, "Cross-language plagiarism detection technology based on fingerprint fusion".