A Systematic Review of Multilingual Plagiarism Detection: Approaches and Research Challenges

Chaimaa BOUAINE, Faouzia BENABBOU, Zineb Ellaky, Amine BOUAINE, Chaimae ZAOUI
Laboratory of Artificial Intelligence and Systems, Faculty of Sciences Ben M'Sick, Hassan II University, Casablanca, Morocco

***Abstract*—The existence of voluminous multilingual sources on the web in different fields creates numerous issues, including violations of intellectual property rights. For that, the multilingual plagiarism or cross-language plagiarism detection (CLPD) has become a great challenge, which refers to copying content from a source text in one language into a target text in another without proper attribution. This study presents a systematic literature review (SLR) of methodologies used in CLPD covering works published between 2014 and 2025. This literature review summarizes and diagrams the different approaches used for CLPD. We propose a classification of the different representations of multilingual texts into four types: traditional approaches, multilingual semantic networks, fingerprinting methods, and deep learning models. In addition, we have carried out an in-depth analysis of ten language pairs, have focused on the approaches employed, including translation strategies, feature extraction approaches, classification techniques, similarity methods, dataset types, data granularity, and evaluation metrics. Among the fulfilled results, English appears in 98% of language pairs, and the English-Arabic pair stands out as the most studied. Over 60% of studies involve a translation phase with Google Translate as the most frequently used tool. The mBART model achieves over 95% accuracy for English-Spanish, English-French, and English-German, while BERT reached 96% for English-Russian. As for the assisted translation study based on the Expert translation tool, strong results are obtained for English-Persian, with an accuracy of 98.82%. On the whole, transformers offer better results in several language pairs without the need for translation.**

Keywords*—*Multilingual plagiarism; systematic literature review; multilingual text representation; translation approaches; natural language processing; machine learning; deep learning

I. Introduction

Easy access to scientific and, more specifically, academic documents is a major asset for academic research. However, it is a double-edged sword, as it has led to an increase in intellectual property infringements: users do not always mention source documents. Today, in a multilingual environment, the rich existence of translation tools has further increased this problem and encouraged the growth of CLPD, where content is copied from one language and translated into another. Plagiarism is broadly defined as the unauthorized use or imitation of another person's work, ideas, or expressions, presented as if they were one's own, without attribution [1],[2]. It can take many forms, including plagiarism of ideas [3], verbatim copying and pasting [4], self-plagiarism [5], paraphrasing [6], and translation [7]. CLPD particularly refers to the act of translating content from one language to another, frequently involving paraphrasing or altering the format to mask the source [8]. There are two main types of plagiarism detection techniques: intrinsic and extrinsic. To find potential parallels, extrinsic approaches evaluate a suspect text by comparing it with information from other sources, such as databases or published documents [9],[10]. This technique works well for identifying cut-and-paste, translations, and basic paraphrases, but its effectiveness depends on the caliber and scope of the reference material used. Intrinsic methods focus on internal components such as lexical diversity, sentence structure, and writing style. Without using an external corpus, they can identify stylistic irregularities that may be signs of plagiarism [11], [12].

The CLPD system presents significantly more complex challenges than monolingual plagiarism detection due to the linguistic and cultural diversity between source and target texts. Firstly, the language barrier makes it difficult to directly identify similarities when texts are written in different languages [13]. Secondly, reliance on machine translation systems can introduce semantic or syntactic errors and compromises the reliability of post-translation comparisons [14]. Moreover, authors intending to conceal plagiarism may reformulate ideas using stylistic variations [15] and paraphrasing specific to the target language [16], making it even harder to detect correspondences. Finally, in many cases, there are no aligned parallel resources or sufficient bilingual corpora for certain language pairs, which limits the effectiveness of approaches based on comparable or aligned corpora [17].

To meet these challenges, CLPD relies on several main approaches. The first is to use machine translation to convert documents into a common language, and then apply traditional text-matching methods [18]. Although simple, this strategy is highly dependent on the quality of the translation. A second approach avoids direct translation by using multilingual embedding models that project texts from different languages into a shared vector space, enabling direct semantic comparison [19]. Another method uses knowledge graphs, where multilingual semantic networks such as WordNet are employed to represent conceptual relationships between terms [20]. Finally, the fingerprinting technique extracts distinctive features (such as n-grams) from texts and compares them across languages [21]. These approaches make it possible to identify semantic similarities without relying solely on word-for-word translation.

This study aims to conduct an SLR of CLPD. To the best of our knowledge, only one article has addressed multilingual text plagiarism detection [22]. It is crucial to conduct a review of the existing reviews on CLPD because the field of CLPD detection is constantly evolving, and new methods and techniques are being developed. This SLR covers research on CLPD techniques published between 2014 and 2025 and addresses

various Multilingual Text Representation Strategies. This SLR study offers a concise and clear summary of the methodology currently used in CLPD. Our work provides a comprehensive analysis of all phases involved in CLPD, including data preprocessing, feature extraction, models used, similarity measures, and dataset utilization. Additionally, we extracted a taxonomy of methodologies employed in CLPD, a level of methodological integration and synthesis that had not been thoroughly addressed in prior state-of-the-art research.

The significant contributions of this research work are as follows:

- Comprehensive review of most studies addressing CLPD.
- Proposition of a systematic classification of multilingual text representations into four main categories: traditional approaches, multilingual semantic networks, fingerprinting methods, and deep learning models.
- In-depth analysis of ten language pairs, emphasizing the approaches employed, including translation strategies, feature extraction methods, classification techniques, similarity measures, dataset types, data granularity, and evaluation metrics.
- Detailed examination of detection techniques specific to each language pair.
- Proposition of an architectural framework for CLPD, integrating the insights gained from the literature and the conducted analyses.

This research study is structured as follows: Section II presents the related work. Section III presents the systematic review methodology. Section IV offers an overview of the state-of-the-art techniques in CLPD, examining the considered language pairs, feature extraction strategies, datasets, and model performances. We present the analysis of the research questions and structure the analysis into two distinct parts for better clarity and understanding. The first part provides a general overview and examines all the approaches used for CLPD. The second part focuses on a detailed and specific analysis of each language pair. In Section V, we generate the architectures employed in CLPD methods and results. And, we highlight promising research directions that remain largely unexplored. Finally, we summarize the main results and shed light on future work in Section VI.

II. Related Work

This section aims to present a summary of studies on CLPD approaches, exploring various approaches and techniques developed between 2014 and 2025. We analyze the studied language pairs, feature extraction strategies, datasets, and model performances. This review highlights the progress made through the integration of advanced models, such as transformers, and the use of multilingual resources while identifying persistent limitations, particularly for under-resourced languages. To structure this analysis, we classify existing works into four main categories based on the feature extraction step: 1) traditional approaches, encompassing classical word embedding techniques such as Word2Vec, TF-IDF, as well as statistical methods like LSI (Latent Semantic Indexing) and SVD, which rely on algebraic transformations; 2) transformers & deep learning, including modern neural-based models such as transformers like BERT, XLM-R, as well as deep learning architectures like GRU, which learn contextualized text representations through neural networks; 3) approaches based on Multilingual Semantic Networks (MSN) such as WordNet, BabelNet, which rely on predefined linguistic knowledge rather than statistical or neural-based learning, often combined with similarity measures such as Wu-Palmer (WuP), Lin, and Jaccard for Cross-Lingual (CL) text similarity assessment; and 4) fingerprinting techniques, which use text fingerprinting techniques like Winnowing, N-grams, and hash-based techniques to capture distinct textual patterns and structures for plagiarism detection and text similarity assessment. A systematic comparison of previous research is easier due to this classification, which clearly distinguishes various methods according to how they represent text and extract features.

A. Traditional Approaches

The authors of this paper [23] proposed a method named CL-WE-Tw, which combines word2vec with part-of-speech (POS) features, the Term Frequency-Inverse Document Frequency (TF-IDF) weighting technique, and MUSE for the Arabic-English (Ar-En) language pair. The proposed model achieved a Pearson correlation (PC) of 81.47% on the SemEval-2017 dataset. In [24], the authors presented an approach for Ar-En CLPD using several techniques, including CL Conceptual Thesaurus-based Similarity Continuous Bag-of-Words (CL-CTS-CBOW), CL Word Embedding Similarity (CL-WES), and Inverse Document Frequency (IDF). The IDF method achieved an F-score of 88% at the word level and 82.75% at the sentence level on four datasets (Books, Wikipedia, EAPCOUNT, and MultiUN). A method to measure semantic textual similarity for Spanish and English (Es-En) sentences was proposed in [25] to detect CLPD. The techniques include CL-CnG (character n-grams), CL-CTS (conceptual thesaurus), CL-WES (word embeddings), and T+WA (word alignment after translation), as well as supervised and unsupervised combinations, notably using the M5' model. The method achieved a correlation of 88.02% on SemEval-2016 and 83.02% on the SNLI corpus of SemEval-2017. In [26], the authors proposed the CL Word Mover's Distance (WMD) method and evaluated similarity in an aligned multilingual space for a Chinese-English text. Vector word representations are formed independently for each language using the Skip-Gram model and aligned using a small bilingual dictionary. On the NDLTD dataset, CL-WMD achieved a Hit@10 of 97.09% at the paragraph level and 86.09% at the sentence level. In [27], a CLPD was developed for the English and Russian languages. The method translates Russian documents into English using a Transformer-based machine translation system and applies semantic clustering with FastText anchors for source retrieval and unsupervised and semi-supervised sentence anchors for document comparison. Using a dataset synthesized from Russian and English Wikipedia, the system achieved an F1 score of 80%. On the PAN'11 dataset, the system achieved a precision of 94%, a recall of 76%, an F1 score of 84%, and a PlagDet score of 83% for monolingual plagiarism detection. The article [28] focused on detecting English-Arabic (En-Ar) using methods to extract semantic and syntactic features, such as word order, word embeddings (Word2Vec),

TF-IDF, and word alignments with multilingual encoders MUSE. These techniques, combined with the different ML algorithms, including Support Vector Classifier (SVC), Logistic Regression (LR), Linear SVC, DT, KNN, and Extreme Gradient Boost (XGBoost), were used to determine whether the sentences were plagiarized. The approach achieved an F1-score of 0.879 on the SemEval-2017 dataset, with the SVC classifier by integrating all the proposed techniques. Authors of this paper [29] suggested a CLPD technique for English-French (En-Fr) utilizing word embeddings generated through the CBOW model alongside Cosine Similarity (CS) to assess both semantic and syntactic similarities. Among the methods employed are CL Word Embedding Similarity (CL-WES), which evaluates Sentence Embeddings (SE) directly, syntactically enhanced embeddings (CL-WESS), and integration strategies such as decision tree-based methods. The evaluation was conducted on a comprehensive dataset combining texts from Wikipedia, conference papers, product reviews, Europarl, and JRC. The approach achieved an F1 score of 89.15% at the chunk level and 88.5% at the sentence level using decision tree fusion. The authors of this article [30] proposed to detect plagiarism between Ar-En pairs using a Semantic Textual Similarity (STS) and Skip-Gram and CBOW embedding techniques. Three learning approaches were implemented: Parallel Mode, Word by Word Alignment Mode, and Random Shuffling Mode. The combination of the Random Shuffling method and the Skip-Gram technique achieved an important performance, with a correlation rate of 75.7% on the SemEval-2017 dataset. In [31], the authors introduced a CLPD approach based on multilingual word embeddings to align plagiarism fragments across languages. The method focused on the German-English (De-En) and Es-En language pairs and is evaluated on the PAN-PC-11 and PAN-PC-12 datasets. For the candidate retrieval phase, potential fragments are identified using a vector space model with TF-IDF-like weighting and CS. For the detailed analysis phase, a word-graph representation is used to capture semantic and syntactic relationships through clique-based graph matching. The method achieved PlagDet scores of 85.7 (De-En) and 83.5 (Es-En) on PAN-PC-11, and 86.2 (De-En) and 84.2 (Es-En) on PAN-PC-12. Additionally, the method achieved a PC of 44.3 on the SemEval 2017 dataset for Es-En. The authors of [32] focused on the candidate retrieval phase for the De-En language pair, aiming to identify potential source documents for suspicious texts. It uses an approach based on thematic segmentation, Latent Dirichlet Allocation (LDA), proximity-based language models, and the extraction of keywords using TF-IDF, along with representative phrases (bigrams and trigrams). Representative words and phrases are translated using Google Translate, reducing reliance on full document translation. Experiments on the PAN-PC-12 corpus show an F2-score of 67.03% by combining thematic and linguistic segmentation. This article [33] focused on candidate document retrieval to effectively identify potentially plagiarized sections between Chinese and English texts. The developed techniques included methods based on keywords using TF-IDF and machine translation, utilizing tools such as BABYLON and Google Translate. Performance results show that queries based on 50% of the keywords provide a balance between efficiency and accuracy (MAP = 0.500) when applied to the Xinhua Chinese and English news collections. Authors of [34] presented an approach based on two steps: candidate fragment identification and detailed analysis. In the first stage, the aim is to identify potentially plagiarized fragments. It begins by selecting a subset of words, termed representative terms, which characterize the vocabulary used in the source document using the TF-IDF method. This approach does not require a complete translation of the text; only the representative terms are translated. Several translation resources were evaluated for this purpose, including Google Translation, BabelNet, and Dict.cc. Among these resources, Google Translation achieved the PlagDet score, reaching 75.42% similarity, with a precision of 72.67%. The second stage involves aligning the source and suspicious fragments, followed by testing their similarity using dynamic programming algorithms. Different term weighting models were utilized, including the TF-IDF model, the Bernoulli weighting model, the Bose-Einstein weighting model, the Simple IDF model, and the Binary model. The TF-IDF model achieved a plagdet of 91.66%, the Bernoulli weighting model reached a plagdet of 91.21%, the Bose-Einstein weighting model obtained a plagdet of 91.88%, the Simple IDF model reached a plagdet of 91.82%, and finally, the Binary model achieved a plagdet of 92.08% specifically for Es-En text pairs. In this article [35], the authors developed two approaches for the candidate retrieval task in CLPD methods. In the first approach, the most representative words for search queries are extracted using the TF-IDF technique, while in the second, the concepts are extracted and documents are semantically represented using the CL Explicit Semantic Analysis (ESA) method. The CL-ESA model was developed using Wikipedia by utilizing interlingual relationships between English, German, and Spanish articles. This allowed content to be mapped into a common conceptual space to handle problems such as polysemy and synonymy. These two methods were merged to create a hybrid model that uses dynamic interpolation to integrate similarity scores based on concepts and keywords. Evaluated on datasets such as PAN-PC-12, JRC-Acquis, PAN-PC-11, and Wikipedia, the hybrid model achieved a recall of 78.89%, a precision of 61.74%, an F1 score of 69.27%, and an F2 score of 74.74. In [36], the authors proposed a CLPD method for the English-Persian language pair, and five methods were used, each with distinct techniques and precision levels. The Translation plus Mono-lingual Analysis (T+MA) method, combining machine translation and mono-lingual analysis, achieved the highest precision of 83%. The Bilingual Word Embeddings Without Alignment (BILBOWA) method, using bilingual word embeddings, reached a precision of 55%. CL Latent Semantic Indexing (CL-LSI) created a multilingual semantic space and achieved 49% precision, while CL-ESA, based on similarity to Wikipedia documents, resulted in a lower precision of 21%. Lastly, CL Knowledge Graph Analysis (CL-KGA), utilizing BabelNet for knowledge graph analysis, showed an intermediate precision of 70%. The article [37] proposed a method for detecting bilingual plagiarism for En-Pe documents. Based on the Vector Space Model (VSM), the technique utilizes morphological analysis, synonym lists, and bilingual dictionaries to compare textual content. The approach relies on term weighting with TF-IDF to assign a numerical weight to each word. The Text similarity is then measured using CS. The dataset includes 100 training texts and 100 test texts extracted from internet sources. The performance is evaluated with 88% precision, 96% recall, and an F-measure of 91%. In

[38], the authors addressed the identification of Vietnamese-English paraphrases using Siamese recurrent architectures, incorporating techniques such as Siamese Long Short-Term Memory (SLSTM), Word2Vec-based bilingual word embedding mapping, adding POS vectors to word embeddings, and revising POS labeling tags using WordNet and VietNet. The proposed method achieved an accuracy of 89.61% on the TED dataset. In this article [39], the system proposed aims to assess semantic similarities between Indonesian and English documents by combining Latent Semantic Analysis (LSA), Singular Value Decomposition (SVD), and Learning Vector Quantization (LVQ) for classification. The system utilizes a local dictionary database for word-by-word translation, ignoring grammatical rules, and calculates similarities through CS using the Slice and Pad methods as well as the Frobenius norm. When tested on manually translated scientific articles, the system achieved a maximum accuracy of 87%. Furthermore, using a term-document matrix based on term frequency significantly outperforms a binary method in terms of accuracy. The authors in [40] addressed the challenge of En-Ar texts by using latent semantic indexing (LSI) for both English and Arabic texts. Latent semantic indexing constructs a shared semantic space for both languages, enabling contextual similarity comparisons between documents without the need for direct translation. When utilized on the parallel En-Ar Corpus of United Nations documents (EAPCOUNT), the LSI technique resulted in a similarity detection rate of 93%, while Jaccard's index yielded just 33%. The authors of this article [41] developed a CLPD system that examines both syntactic and contextual similarities among texts in various languages, such as English-Chinese, English-Japanese, and English-Korean. The system employs sophisticated methods, including Word2Vec and a model based on Convolutional Neural Networks (RCNN). It achieved commendable accuracy on the student dataset, reporting 88.87% for English-Chinese, 87.49% for English-Japanese, and 87.26% for English-Korean. In this article [42], a system is proposed for detecting text reuse for the English-Urdu languages through the development of the CLEU corpus. The goal is to create a realistic resource for detecting multilingual plagiarism. The method used, Translation Plus Monolingual Analysis (T+MA), first translate the Urdu text into English using Google Translate, before applying similarity techniques such as n-gram overlap, Longest Common Subsequence (LCS), and Greedy String Tiling (GST). Pairs of sentences/passages are classified as Near Copy, Paraphrased Copy, and Independently Written. The results show an F1 score of 73.2% in binary classification and 55.2% in ternary classification.

B. Multilingual Semantic Network

A Multilingual Semantic Network is a knowledge base that represents semantic relationships between words or concepts across multiple languages. It captures the meanings and relationships between terms in different languages, enabling tasks such as machine translation, multilingual information retrieval, or natural language understanding (NLU) in a multilingual context. These semantic networks are often built by linking words or concepts across languages through lexical or conceptual alignments, using resources like WordNet [43], BabelNet [44], or other multilingual lexical databases.

Fuzzy semantic similarity techniques were developed in [45] for Ar-En within a Big Data framework. Performance metrics were based on the Wu&Palmer and Lin similarity measures, leveraging WordNet to assess semantic relationships between words. Experiments conducted in a Hadoop environment (HDFS and MapReduce) demonstrated that the Wu&Palmer measure achieved a precision of 54%, recall of 66%, and F-measure of 59.4%, while the Lin method achieved a precision of 27%, recall of 37%, and F-measure of 31.2% on a dataset comprising news, articles, tweets, and academic works. The work in [46] addressed CLPD for the Urdu-English language pair using ML models. Using Jaccard and CS were calculated for uni-grams and tri-grams, with lemmatization performed using WordNet during preprocessing. Five classifiers were utilized: Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF). Among them, KNN achieved an accuracy of 92%. In the article [47], the authors proposed the CL Ontology-Based Similarity Analysis (CL-OSA) model, focusing on multilingual text similarity for languages such as Es-En, De-En, Chinese-Japanese, and French-English. The CL-OSA model leverages Wikidata, a multilingual open knowledge graph, to represent documents as language-independent entity vectors. CL-OSA applies a semantic hierarchy-based weighting and CS to rank document similarities. Evaluation of datasets such as PAN-PC-11, ASPEC-JE, ASPEC-JC, JRC-Acquis, and Europarl. For candidate retrieval, CL-OSA achieved a Mean Reciprocal Rank (MRR) of 91.38% for Es-En, 71.92% for Japanese-English, 78.21% for Japanese-Chinese, 97.68% for French-English, and 55.47% for French-English. For detailed analysis, CL-OSA achieved a PlagDet score of 57.3% for Es-En and 52.1% for De-En. The authors in [48] proposed a Multi-Lingual Plagiarism Detection (MLPD) model for English-Persian texts, focusing on translational plagiarism. The method utilizes translation tools, including manual translations by English experts and Google Translate, to translate Persian texts into English, followed by semantic similarity evaluation using the WUP metric. In the candidate retrieval phase, the system uses Apache Solr to identify the five most probable source documents by calculating the frequency of suspicious phrases in the indexed dataset. The findings highlight the significance of accurate translations on the Mizan dataset, with the suggested method achieving an accuracy of 98.82% with expert translations and 56.9% with Google Translate. A CL Knowledge Graph Analysis (CL-KGA) method was used in [49], which uses BabelNet to construct knowledge graphs and measure semantic similarity between document fragments. The languages studied are De-En and Es-En. The performances show that for automatic translations (AT), the PlagDet score reaches 60.87% for the Es-En pair and 52.96% for the De-En pair, while paraphrastic translations (PT) reach scores of 9.93% and 10.06%, respectively. The similarity measures are based on intersection algorithms and the weighting of concepts and graphical relationships. In [50], the authors proposed a method for detecting similarities between Arabic and English documents using Linear LR for classification. The approach starts with key phrase extraction and the translation of Arabic documents into English, followed by similarity detection using techniques like CS, LCS, and N-grams (tri-grams), with the integration of synonyms through WordNet. Tested on Wikipedia articles, the method achieved a precision of 96%, a

recall of 85%, and an F-measure of 90%. This work [51] proposed a fuzzy approach to inter-language (French-Arabic, En-Ar) plagiarism detection by combining semantic similarity with an Apache Hadoop-based Big Data environment. Documents are pre-processed (tokenization, empty word removal, POS tagging, and trigram segmentation) and analyzed using WordNet and similarity measures such as WuP, Lin, and Leacock-Chodorow (LCH). A test corpus of 600 documents was used, including 200 automatically translated and 400 translated with modifications (paraphrasing, back-translation). The results show that WuP delivers the best performance, with a similarity percentage of around 63%. In [52], the authors addressed two tasks: CL Text Semantic Similarity (CL-STS) and plagiarism detection (PD) for En-Ar texts. It relies on several semantic features, including topic similarity, Named Entity Recognition (NER), semantic role labeling (SRL), spatial role labeling (SpRL), stop word bag, and meaning bag. Topic generation is performed using LDA, complemented by the use of BabelNet to extract English synonyms from Arabic topics and the WUP metric to assess semantic similarity. Arabic texts are translated into English using the Google Translate API to enable analysis in a common language. For plagiarism detection, three models were used: Deep Neural Networks (DNN), LR, and SVM. The results showed that the SVM model achieved an accuracy of 96.65%, the LR model 96.64%, and the DNN model 97.01%.

C. Fingerprints

In this study [53], the authors developed a method for detecting CLPD for En-Ar, utilizing techniques such as keyphrase extraction to compute phrase frequency and rank candidate keyphrases, along with C-Value and NC-Value algorithms, translation of keyphrases into English, and fingerprinting for document representation. Five similarity measures were employed: N-Grams Similarity, LCS, Dice Coefficient (DC), Jaccard Similarity, and Containment Similarity, which were used as features to train three ML models: SVM, NB, and LLR. The results demonstrated that SVM achieved an F-score of 92% when more than three similarity measures were combined. This paper [54] proposed a method for Chinese-English based on WordNet. The approach extracts nouns using ICTCLAS for Chinese and the Stanford POS Tagger for English, encoding them into language-independent fingerprints via WordNet's semantic hierarchy. A disambiguation algorithm based on semantic density calculates the relevance of senses to address word polysemy. The method measures similarity using the Dice coefficient to compare fingerprints and identifies potential plagiarism cases when the similarity exceeds a predefined threshold. Tested on a parallel corpus, National Knowledge Infrastructure (CNKI), the proposed method achieved a precision of 87% and a recall of 78%. In [55], the authors developed a plagiarism detection system for multilingual documents (English-Indonesian) using the Winnowing method and the Jaccard coefficient to measure similarity. The system allows users to upload documents in text or PDF format, translate them automatically, if necessary, pre-process them (folding capital letters, tokenization, removal of empty words), and generate fingerprints for comparison. The dataset used includes academic journals from sources such as IEEE Xplore and ResearchGate. The system achieved an accuracy of 84.7%. The authors of [56] proposed for the Ar-En language pair, the method utilizes the Winnowing algorithm to generate digital fingerprints from k-grams extracted from documents, achieving 81% recall, 97% precision, and an 89% F-measure on Wikipedia articles.

D. Transformers & Deep Learning

The pre-trained models Multilingual BERT (M-BERT) and CL Roberta (XLM-R) [57] were used for the English-Vietnamese language pair. For this task, XLM-R achieved an accuracy of 84.3% and an F1-score of 87.6%, while M-BERT achieved an accuracy of 73.7% and an F1-score of 81.3% on the GLUE dataset. An AraXLM method based on the XLM-RoBERTa model was proposed in [58] to detect plagiarism for Ar-En texts. The framework utilizes the SemEval-2017 Task1 dataset, where the Arabic sentences were translated from English using Google Translate. The approach includes automatic diacritization, semantic similarity calculation using FAISS (Facebook AI Similarity Search), and measures such as CS and Euclidean Distance (ED) to compare the embeddings of sentences. In testing, CS achieved 94.49%, and ED was measured at 1.74668 for non-diacritized sentences. This paper [59] proposed an mBART transformer for feature extraction combined with SLSTM. Experiments were conducted on language pairs including En-Fr, En-Es, and En-De. The methodology achieved an accuracy of 98.83% and an F1-score of 98.87% for En-Fr, 97.94% accuracy and an F1-score of 98.01% for En-Es, and 95.59% accuracy with an F1-score of 96.02% for En-De. The evaluation was performed on datasets such as PAN-11, JRC-Acquis, Europarl, Wikipedia, and conference papers. This work [60] explored various methods for addressing Russian-English text alignment in the context of plagiarism detection. The method translated the Russian text into English using Neural Machine Translation (NMT). To identify translated plagiarism, a comparative analysis of models including SE, BERT, Word Substitution (WS), and LASER was conducted. With an F-score of 95%, a precision of 96%, and a recall of 93%, the BERT model stood out among the others. The study also assessed the similarity of translated sentences using Jaccard metrics with 1-grams (NMT) and 2-grams (NMT2). The NMT model demonstrated a precision of 85%, a recall of 80%, and an F-score of 82%. Furthermore, LR was applied in two configurations: LR-1, which incorporated all techniques, and LR-2, which focused on SE and WS. LR-1 attained a precision of 91%, a recall of 80%, and an F-score of 85%. In [61], the authors addressed the case of plagiarism in the Persian-English pair using multilingual transformers models (XLM-R, M-Bert, DistilBert) and the CS metric. The findings of the study indicate that the XLM-RoBERTa model achieved a PC of 95.62% on the PESTS dataset. In comparison, M-BERT achieved a correlation of 91.88%, while DistilBERT achieved a correlation of 89.51%. The authors of [62] proposed a technique that combines data augmentation with a vigilant Siamese LSTM model to detect plagiarism between Tibetan and Chinese. For Tibetan-Chinese sentence pairings produced from the CWMT and SICK datasets, this method produced a PC of 54.76%. The similarity between document abstracts was also assessed using a Doc2Vec model, yielding Tibetan-Chinese PC scores of 87.06% for Chinese, 63.67% for Tibetan, and 53.9% for Tibetan-Chinese. This work [63] proposed a CLPD approach using a graph transformer-based model (CL-GTA) and knowledge graphs (KG). After creating KG from the Extended Open Multilingual WordNet, the model employed a graph transformer to weight entities and

semantic relations using a multi-head attention mechanism. CS was applied to measure text similarity, and the system was evaluated on multiple datasets, including Europarl, JRC-Acquis, Wikipedia, PAN 2011, and Ar-En parallel corpora. The CL-GTA model achieved a PlagDet of 62% for Es-En, 58.4% for French-English, and 52.2% for Ar-En**.** In [64], the authors proposed Doc2Vec+SLSTM CLPD in the Es-En pair combines Doc2Vec for text representation and an SLSTM model to evaluate document similarity. The performances were compared to other techniques, such as GloVe, FastText, BERT, Word2Vec, and Sent2Vec, using datasets from PAN11, JRC-Acquis, Europarl, and Wikipedia. The Doc2Vec+SLSTM model achieved the highest accuracy of 99.81%, outperforming GloVe (99.59%), FastText (98.82%), BERT (99.49%), Word2Vec (99.14%), and Sent2Vec (98.41%). This paper [65] identified paraphrases for English and Vietnamese sentences using a hybrid approach. The method integrates a Fuzzy-based approach leveraging BabelNet to evaluate semantic relationships between words, a Siamese LSTM model to calculate sentence similarities, and feature combination methods employing algorithms such as LR, RF, and Multilayer Perceptron (MP). BabelNet, enhanced with VietNet, addressed its limitations by adding more Vietnamese words to the synsets. The model was evaluated on a bilingual English-Vietnamese corpus derived from TED. The proposed approach with Linear Regression yields a precision of 80.3%, a recall of 95.8%, and an F-measure of 87.4%. In [66], the authors evaluated the effectiveness of six multilingual transformer-based models for CLPD. The models tested are mBERT, mDistilBERT, XLM-RoBERTa, SBERT Multilingual MiniLM-L12, SBERT Multilingual MPNet, and Distil SBERT Multilingual. The study was conducted on language pairs formed between English and ten languages from the Indo-European family: En-Es, En-Pt (Portuguese), En-Fr, En-Ru, En-Sr (Serbian), En-Cs (Czech), En-De, En-Nl (Dutch), En-Sv (Swedish), and En-Hy (Armenian). The datasets used include sentences from Wikipedia translations to generate positive examples and test sets derived from the Microsoft Research Paraphrase Corpus (MRPC), as well as specific sets like Negative-1 and Negative-4 for the En-Ru pair. The authors of [67] presented an approach focused on the English-Urdu language pair. It leverages recent multilingual models such as LLaMA and Mistral to extract shared semantic representations across languages. These representations are then used to compare content between different languages. The evaluation was conducted on the CLPD-UE-19 dataset, and the results report an F1 score of 73.9%.

III. Research Methodology

A systematic literature review is a structured method of reviewing academic literature that involves collecting and critically evaluating a set of articles focused on a specific topic. Its purpose is to identify, select, synthesize, and analyze relationships, limitations, and key findings, thereby providing a comprehensive summary of both quantitative and qualitative studies [68]. Conducting an SLR is essential for researchers to explore current research trends related to the detection techniques of Multilingual Plagiarism, and to identify weaknesses in existing methodologies.

A. Search Strategy

To ensure the relevance and comprehensiveness of the studies retrieved, we make use of the most widely recognized research databases, as the choice of search strategy significantly influences the quality and completeness of the results. The databases included Scopus, ScienceDirect, IEEE Xplore Digital Library, SpringerLink, ACM Digital Library, ResearchGate, Web of Science, arXiv, and Google Scholar. Additionally, the Rabbit tool is employed to enhance the efficiency of the search process and to enable the identification and organization of relevant studies more effectively.

B. Research Questions

To gain a comprehensive understanding of the issue of CLPD along with its associated challenges, we formulate the research questions that this SLR aims to address as follows:

RQ1: Which language pairs are most commonly studied in CLPD, and which languages dominate research?

RQ2: What are the common translation strategies used in CLPD systems, and what are the main types of approaches adopted in this field?

RQ3: What are the different multilingual representation methods used in the feature extraction step, and what is their impact in terms of performance?

RQ4: What similarity measures are used in CLPD, and which one is the most commonly dominant in the existing studies?

RQ5: What are the main factors (preprocessing, datasets, multilingual semantic networks, feature extraction techniques, ML/DL models, Similarity Measures) that influence performance in each language pair?

C. Query Terms

Our research utilizes a "snowball" method to identify relevant literature in the field of CLPD, and starts the literature review with a few critical or foundational selected studies. Based on the references included in these selected studies, additional publications are found. This process is carried out repeatedly. The approach persists until no additional pertinent studies are discovered. The detailed steps of our snowball search method are outlined below.

- Applying text mining and natural language processing to an initial set of CLPD studies enabled the extraction of key terms, including concepts like cross-language alignment, semantic similarity, and translation-based detection.
- Constructed from these terms, a search query using relevant keywords and Boolean operators is formulated to ensure effective retrieval.
- Conducted across major databases such as IEEE Xplore, SpringerLink, ScienceDirect, ACM Digital Library, and Google Scholar, the search aimed to capture both breadth and depth.

- Manually reviewed, the results were categorized as Relevant (R) or Not Relevant (NR) based on their alignment with CLPD objectives.
- Refined through further analysis, the keyword set is adjusted with generic terms removed and domain-specific ones like “semantic similarity” or “sentence alignment” included.
- Extended by snowballing, each relevant paper’s references and subsequent citations are explored to trace both foundational and recent contributions.
- Repeated iteratively, the process concludes when no additional relevant studies emerge, resulting in a curated set of literature for the CLPD review.

Query 1: ("cross-language plagiarism" OR "cross-lingual plagiarism" OR "multilingual plagiarism detection" OR "cross-language text reuse" OR "translation plagiarism").

Query 2: ("plagiarism detection" OR "copy detection" OR "text similarity detection" OR "text matching" OR "semantic similarity").

Query 3: ("multilingual embeddings" OR "word embedding" OR "sentence embedding" OR "machine learning" OR "deep learning" OR "transformers").

Query 4: ("translation-based" OR "translation independent" OR "machine translation" OR "neural translation").

Query 5: ("systematic literature review" OR "SLR" OR "survey" OR "comparative analysis" OR "performance evaluation").

D. Study Selection

To ensure the transparency and scientific rigor of our SLR, we defined explicit inclusion and exclusion criteria to guide the selection process. After retrieving the initial set of studies through our search queries, a two-stage screening was conducted. In the first stage, titles and abstracts were reviewed to quickly discard irrelevant works. In the second stage, the full texts of the remaining studies were assessed against the predefined criteria. Only those studies that satisfied all inclusion criteria and none of the exclusion conditions were retained. This systematic filtering ensured that the final corpus focused exclusively on peer-reviewed, English-language research directly addressing CLPD with robust methodological contributions. The complete list of inclusion and exclusion criteria is presented in Table I.

TABLE I. LIST OF INCLUSION AND EXCLUSION CRITERIA

Inclusion Criteria	Exclusion Criteria
Articles between 2014 and 2025	No experimental methods are employed.
The papers are written in English	Papers unrelated to the research questions defined in our SLR.
Scientific papers are published through conferences or journals.	Closed access or missing full text.
The study focuses on CLPD	Research papers use monolingual plagiarism
The articles apply one or more of the following methods: MSN, feature extraction, machine learning, or deep learning.	Preprint papers

E. Quality Assessment

The quality assessment stage in an SLR aims to rigorously evaluate the methodological validity, reliability, and overall relevance of the studies selected [69]. This step acts as a decisive filter, applied to the full-text articles, and marks the final stage in preparing the dataset for data extraction and synthesis. To ensure the inclusion of only high-quality and meaningful contributions, a multi-phase evaluation process is employed [70]. The first phase consists of a preliminary screening of the title, abstract, conclusion, and keywords to eliminate clearly irrelevant studies. In the second phase, the remaining papers undergo a thorough assessment based on predefined quality criteria, such as methodological rigor, clarity of results, and alignment with the research questions. Only the studies that satisfy these standards are included in the final synthesis. The quality assessment rules are defined in Table II:

TABLE II. LIST OF QUALITY ASSESSMENT QUESTIONS

ID	Quality Assessment Questions
QA1	Does the article explicitly address one or more of our research questions?
QA2	Is the scope of the research clearly outlined and focused on CLPD?
QA3	Is the contribution of the study to CLPD clearly stated?
QA4	Is the contribution well-supported by evidence or evaluation?
QA5	Is the dataset used clearly identified and mentioned?
QA6	Are the dataset’s characteristic language pairs and its level of granularity (e.g., sentence, paragraph, document) adequately described?
QA7	Does the study describe the full process used for CLPD?
QA8	Is the feature extraction phase clearly defined?
QA9	Does the study present a clear and well-structured experiment?
QA10	Are the obtained results properly interpreted and discussed?

F. Systematic Literature Review (SLR)

SLR is conducted to identify and analyze relevant research on CLPD techniques published between 2014 and 2025. An initial pool of 1300 articles is retrieved from major scientific databases, including IEEE Xplore, ScienceDirect, Springer, Google Scholar, and Scopus. After an initial screening phase, 57 records are excluded for not meeting basic inclusion criteria, specifically, lack of open access, non-English language, or absence of multilingual representation approaches. The remaining 1243 records are then screened more closely, excluding 1160 articles due to duplication or non-relevance based on title and abstract. This process reduces the selected articles for full-text review to 83. At the next stage, 18 articles are excluded for reasons such as lack of experimental validation, being review papers, or omitting baseline or performance metrics. The final 61 articles are assessed in depth for methodological quality, relevance to the research objective, and overall contribution. Ultimately, 43 studies meet all criteria and are included in the final analysis. The entire selection process and applied filters are illustrated in Fig. 1.

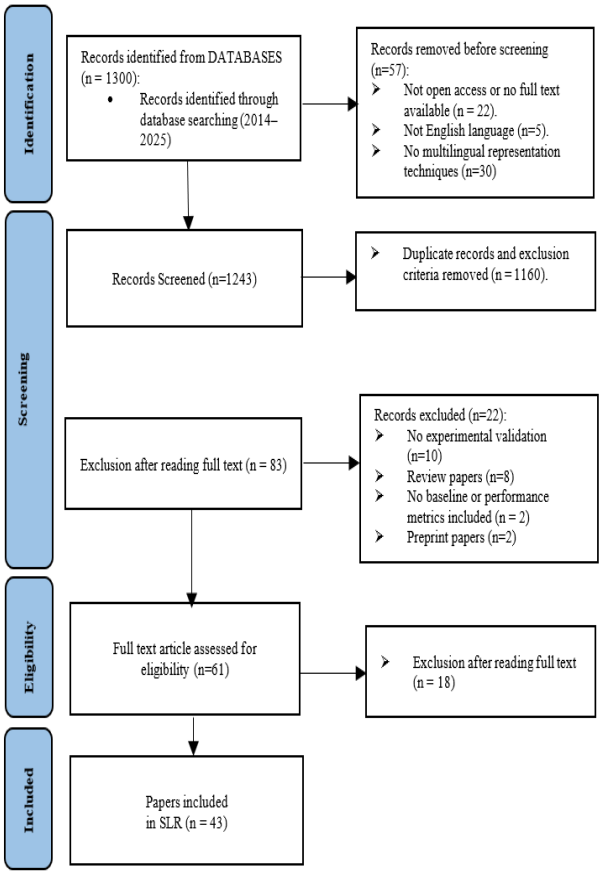


Fig. 1. PRISMA flowchart of the research process.

G. Data Analysis

A total of 43 studies related to the field of CLPD published between 2014 and 2025 are identified. Fig. 2 illustrates the distribution of study types over time. Journal articles dominate the landscape and comprise approximately 57% of the selected studies, followed by conference papers, which represent around 36%, while book chapters constitute the smallest portion, with only 7%. Table III presents the distribution of the retrieved articles across different publication venues. The findings indicate that 59.52% of the studies were published in journals, while 35.71% appeared in conference proceedings. A smaller proportion, 4.76%, was published as book chapters.

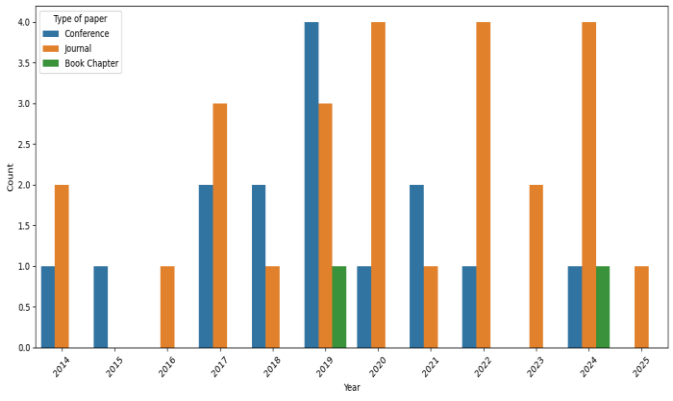


Fig. 2. Distribution of research by type and by year.

TABLE III. DETAILS THE CORRESPONDING PUBLICATION VENUES

Document Type	Publication Title	Reference
Conference Article	Artificial Intelligence XXXVII	[23]
	Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"	[60]
	Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence	[54]
	International Symposium on Communications and Information Technologies (ISCIT)	[38]
	8th SASTech 2014 – Symposium on Advances in Science & Technology	[37]
	Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing	[45]
	Ivannikov Memorial Workshop (IVMEM)	[66]
	International Conference on Developments of E-Systems Engineering (DeSE)	[40]
	Intelligent Computing	[58]
	Proceedings of the Fourth Arabic Natural Language Processing Workshop	[30]
	Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)	[28]
	Conference on Neural Information Processing Systems (NeurIPS 2019)	[27]
	Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents	[47]
	International Workshop on Semantic Evaluation	[25]
	Conference of the European Chapter of the Association for Computational Linguistics	[29]
Journal Article	Information Processing & Management	[32]
	Applied Mechanics and Materials	[33]
	Journal of Information Science	[34]
	Information Processing & Management	[35]
	Intelligent Data Analysis	[36]
	IAES International Journal of Artificial Intelligence (IJ-AI)	[59], [63]
	International Journal of Interactive Mobile Technologies (iJIM)	[64]
	Language Resources and Evaluation	[61]
	Journal of the Association for Information Science and Technology	[42]
	Information Technology Journal	[56]
	The International Journal of Multiphysics	[41]
	Algorithms	[39]
	Journal of Applied Intelligent System	[55]
	Engineering, Technology & Applied Science Research	[57]
	International Journal of Advanced Computer Science and Applications (IJACSA)	[24]
	Asian Journal of Research in Computer Science	[53]

	Proceedings of the Association for Information Science and Technology	[26]
	Data Intelligence	[62]
	Journal of Heuristics	[65]
	Journal of Computing & Biomedical Informatics	[46]
	Journal of AI and Data Mining	[48]
	Journal of King Saud University - Computer and Information Sciences	[52]
	Expert Systems with Applications	[31]
	The European Journal on Artificial Intelligence	[67]
	Journal of Theoretical and Applied Information Technology (JATIT)	[50]
Book Chapter	Recent Advances in Intuitionistic Fuzzy Logic Systems: Theoretical Aspects and Applications	[51]
	Bridging Between Information Retrieval and Databases: Revised Tutorial Lectures of the PROMISE Winter School	[49]

IV. Methods Analysis and Comparison

This section presents the analysis of the research questions, along with the findings derived from the state-of-the-art review. It is structured into two main parts. The first part provides a general overview, examining all the approaches used for CLPD. It highlights the techniques employed, the models applied, and the various language pairs studied. The second part focuses on a detailed and specific analysis of each language pair. For each pair, we evaluate the performance of the employed approaches, considering the linguistic characteristics and datasets used. This in-depth comparative approach highlights the most effective methods for each language pair.

A. Language Pairs (RQ1)

The most commonly used language pairs reveal interesting trends. The En-Ar combination stands out as the most frequently studied, followed by En-Es, En-De, and En-Ar, highlighting the significance of these languages in academic research. In contrast, pairs like Tibetan-Chinese, French-Arabic, and English-Japanese are significantly less explored, due to either a lack of linguistic resources or limited research interest. The omnipresence of English in almost all combinations reflects its dominant position in scientific publications and advancements in natural language processing (NLP). Languages frequently paired with English, such as Arabic, Chinese, or Spanish, are often chosen for their geopolitical importance, widespread use, or linguistic complexity.

It is also relevant to consider the direction of language pairs, for instance, for English-Arabic and Arabic-English, where the source and target languages may differ. This distinction influences the challenges encountered, particularly in terms of machine translation or algorithm adaptation. Furthermore, pairs involving underrepresented languages, such as Tibetan or Persian, often suffer from a lack of annotated corpora and suitable NLP tools, hindering their detailed study. Finally, this analysis reflects not only the current state of research but also the linguistic biases within the field, emphasizing the urgency to diversify efforts to include more marginalized languages and promote linguistic equity on a global scale. Fig. 3 describes the most frequently used language pairs.

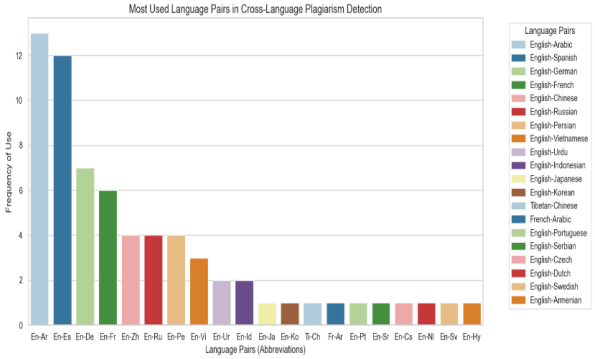


Fig. 3. Frequency of CLPD language Pairs.

B. Translation Approach (RQ2)

In this part, we will examine how translation and non-translation are utilized within the CLPD context, as well as the translation tools employed and the languages being translated.

1) Translation strategies in CLPD: Translation is used in CLPD to convert documents or sentences into the same language, representing a dominant share of 60.5% in the reviewed studies. English is the most studied language, and in most cases, documents are translated into English to ensure both texts are in the same language rather than translated into other languages. Among the translation tools, Google Translate stands out as the most widely utilized due to its accessibility, multilingual support, and continuous improvements in translation quality. However, several studies also explore the use of specialized dictionaries, particularly for domain-specific texts, offering more accurate translations tailored to the subject matter. Additionally, advancements in NLP have led to the adoption of transformer-based models, such as BERT and GPT, which provide context-aware translations and handle complex linguistic structures. Fig. 4 shows the distribution of translation usage.

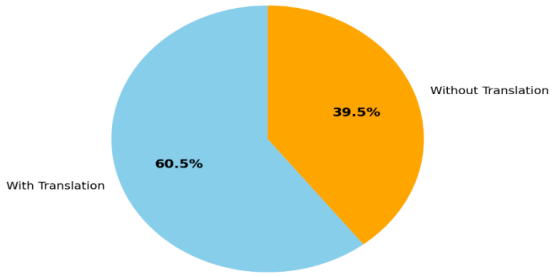


Fig. 4. Distribution of research based on the use of translation.

2) Translation tools used in CLPD: Fig. 5 shows the frequency of translation tools used in CLPD systems, in ascending order of use. Less frequently used tools, such as Transformer, English Expert, BABYLON, and the Dictionary database, appear only once in the studies examined. Tools such as dict. and Bilingual translators are used twice. In contrast, Google Translate stands out as the most widely adopted translation tool, with 21 occurrences, underlining its predominant role and popularity.

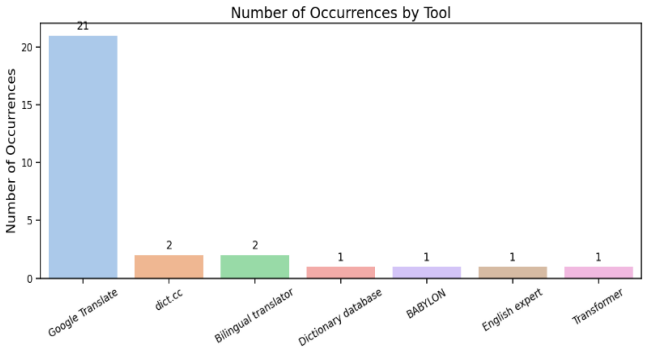


Fig. 5. Frequency of translation tools.

C. Multilingual Text Representation Strategies (RQ3)

This section presents various strategies employed for CLPD, including statistical analysis of approach distribution, a comparison between traditional word embedding methods and transformer-based deep learning models, and the utilization of multilingual semantic networks.

1) Distribution of approaches in CLPD: Fig. 6 highlights the predominance of traditional approaches (47.6%), such as TF-IDF, Word2Vec, LSI, and SVD, which remain widely used due to their simplicity and low computational cost. However, Transformers and Deep Learning models (23.8%) are gaining increasing adoption thanks to their ability to capture rich contextual representations, despite their higher computational requirements. Multilingual Semantic Networks (19%), such as WordNet and BabelNet, play a key role in knowledge-based approaches, particularly for semantic similarity assessment in a multilingual context. Finally, fingerprinting techniques (9.5%), although less common, are employed to capture a unique textual signature.

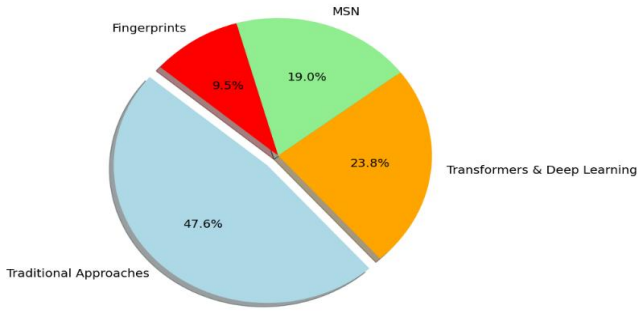


Fig. 6. Distribution of approaches in CLPD.

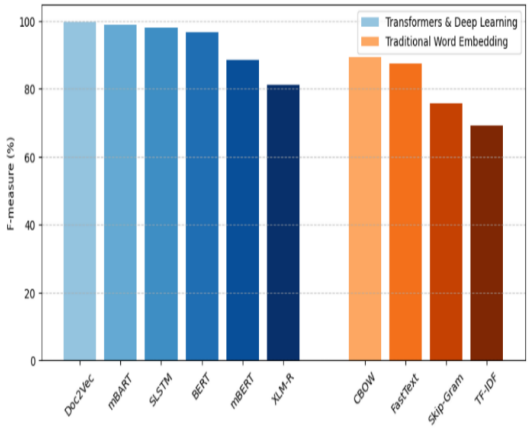
2) Traditional word embedding methods versus transformers & deep learning: Table IV provides a comparative analysis of traditional word embedding methods (e.g., Word2Vec) and transformer-based deep learning models (e.g., BERT, XLM-R) for CLPD. It includes key details such as the studied language pairs, the algorithms employed, the datasets used for evaluation, and the performance metrics. Although various evaluation metrics such as precision and recall are considered, the F1-score is more commonly used as it provides a balanced measure by incorporating both precision and recall.

The F-measure scores, which represent the harmonic mean of precision and recall, are provided to evaluate the effectiveness of these methods. Traditional methods like TF-IDF and Word2Vec often produced results below 80% in terms of F-measure for many language pairs, especially for complex language pairs. For example, TF-IDF/LDA with the En-Es and En-De pairs achieved F-measure scores of 57.58 and 69.27, respectively, while CBOW yielded results of around 78.5. In contrast, Transformer and Deep Learning techniques, such as XLM-R, mBERT, and mBART, surpassed the 80% F-measure mark and demonstrated impressive performance in multilingual contexts, with F-measures ranging from 81 to 98.87, such as 88.5 for En-Vi (XLM-R) and 90.8 for En-De (mBERT). mBART, another transformer model, also performs well, with F-measures of 98.01, 98.87, and 96.02 for the En-Es, En-Fr, and En-De pairs, respectively. In conclusion, transformer-based and deep learning methods generally outperform traditional approaches in terms of F-measure, though results can depend on the specific datasets, and traditional methods remain competitive for simpler tasks or less complex configurations. Fig. 7 compares the performance and temporal evolution of natural language processing models. Fig. 7(a) shows that transformer-based and deep learning models (in blue) generally outperform traditional word embedding methods (in orange) in terms of F-measure, indicating their increased effectiveness. Fig. 7(b) illustrates the timeline of these techniques, revealing that traditional approaches (TF-IDF, Word2Vec) were introduced earlier, while more advanced models like BERT and XLM-R emerged more recently, marking a shift towards more powerful methods over time.

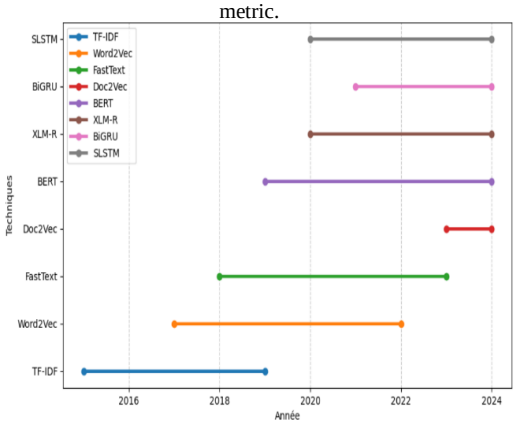
TABLE IV. COMPARISON OF THE PERFORMANCE OF TRADITIONAL WORD EMBEDDING AND TRANSFORMERS & DEEP LEARNING IN CLPD

	ID	Language pair	Algorithm	Dataset	Perf. (%)
Transformers & Deep Learning	[57]	En-Vi	**XLM-R** M-BERT	GLUE	A: 84.3 F1: 88.5
		En-De			A: 87.2 F1: 90.8
		En-Fr			A :86.2 F1 :90.2
	[59]	En-Es	mBART	PAN-PC-11 JRC-ACQUIS EUROPARL Wikipedia Conference papers	A:97.94 P: 98.57 R:97.47 F1:98.01
		En-Fr			A:95.59 P: 95.21 R:96.85 F1 :96.02
		En-De			A: 98.83 P: 98.42 R: 99.32 F1: 98.87
	[60]	En-Ru	BERT	Negative-1	P :96, R :93 F1:95
	[64]	En-Es	Doc2Vec	PAN-PC-11, JRC-ACQUIS EUROPARL, Wikipedia	A:99.81 P:99.75 R:99.88 F1:99.70

	[65]	En-Vi	SLSTM	TED	P :80.3 R :95.8 F1 :87.4
Traditional Word Embedding	[24]	En-Ar	**IDF** CL-WES CBOW	Books, Wikipedia EAPCOUNT MultiUN	**F1:88** F1: 86.5 F1:78.5
	[28]	En-Ar	Word2Vec +TFIDF +MUSE	SemEval- 2017	F1:87.9
	[29]	En-Fr	CBOW	Wikipedia, Conference papers Product reviews, Europarl JRC-Acquis	F1:89.15
	[37]	En-Pe	TF- IDF/VSM	Internet sources	P: 88 R: 96 F1: 91
	[27]	En-Ru	FastText	wikipedia	P: 83 R :79 F1 :80
	[35]	En-Es En-De	TF-IDF	PAN-PC-12, JRC-Acquis PAN-PC-11, Wikipedia	R:78.89 P: 61.74 F1:69.27 F2:74.74
	[32]	En-Es En-De	TF- IDF/LDA	PAN-PC-12	P:46.63, R:75.26 F1:57.58 F2:67.03



(a) Comparison of embedding model performance in terms of the F-measure metric.



(b) Comparison of the embedding model in terms of time evolution.

Fig. 7. Comparison of the performance and temporal evolution of embedding models.

3) Multilingual semantic network: The use of MSN in the field of CLPD serves two main purposes: acting as a dictionary to find synonyms or equivalent concepts across different languages and creating knowledge graphs. According to Fig. 8, we can observe that the majority of research in CLPD does not utilize MSNs, which suggests that many studies rely on alternative approaches, such as techniques based on word embeddings, ML, or DL. However, among the MSNs used, WordNet is the most frequently employed, likely due to its well-established structure and its use as a pivotal lexical resource for several languages. WordNet is often utilized to identify synonyms, antonyms, and other lexical relations, thereby facilitating the detection of semantic similarities. BabelNet, which follows WordNet in usage frequency, stands out for its ability to cover a wide range of languages and integrate information from various sources, such as WordNet and Wikipedia. Other MSNs, such as VietNet, are specifically dedicated to less common languages like Vietnamese, reflecting a focus on local or specific case studies. Finally, Wikidata, a collaborative knowledge base, is less frequently used but offers interesting potential for applications requiring contextualization or encyclopedic information.

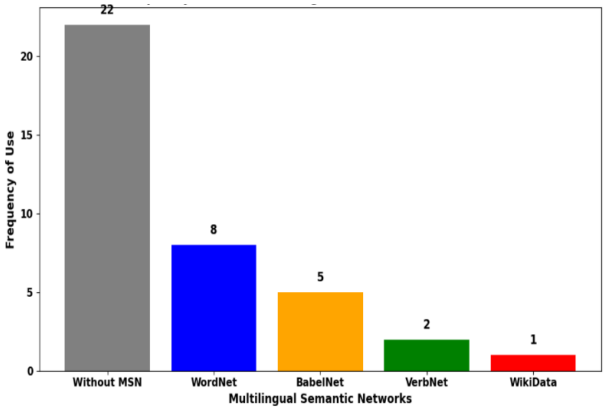


Fig. 8. Frequency of use of multilingual networks.

D. Cross-Lingual Similarity Measures (RQ4)

The most popular similarity metrics in CLPD systems are presented in Fig. 9. Cosine Similarity dominates all studies due to its effectiveness in vector-based representations and its capacity to assess similarities across text lengths. The second most popular index for comparing groups of words or characters is the Jaccard Index. Metrics like WUP similarity, LCS, ED, Lin similarity (Lin), and DC are utilized less often, implying that they are relevant in specific situations or serve as additional methods. Ultimately, the containment measure sees the least use, indicating its specialization in certain scenarios.

E. Comparative Study for each Language Pair (RQ5)

Before analyzing deeply each language pair, the techniques most frequently used in the preprocessing section are discussed, as they are commonly used in the different CLPD approaches.

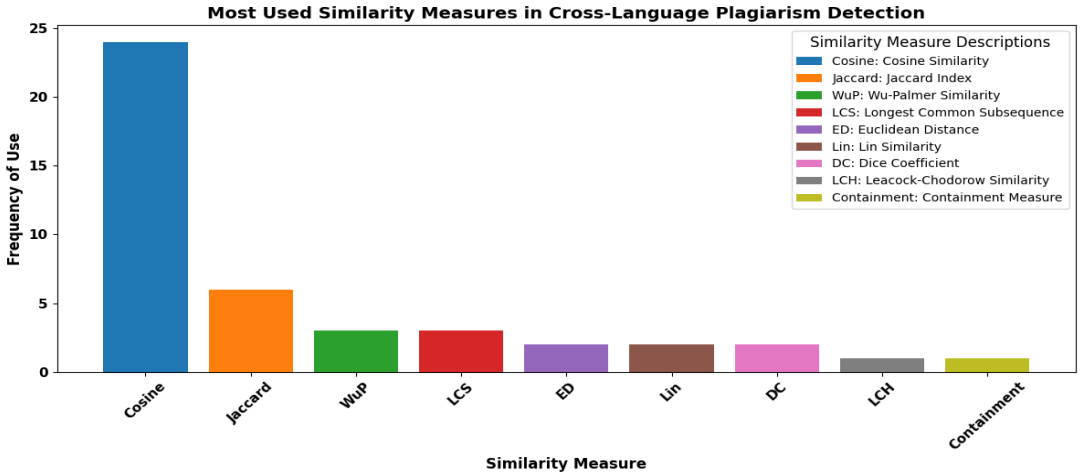


Fig. 9. Frequency of similarity measures in CLPD.

1) Common preprocessing techniques: Text pre-processing in multilingual studies usually begins with the segmentation of the text into sentences or paragraphs, enabling structured and targeted processing. Most of the techniques used at this stage are common and aim to improve data quality. Among the most commonly used steps are tokenization, which divides text into basic units such as words or symbols, and includes converting text to lowercase, removing unnecessary punctuation, lemmatization, and stemming to reduce words to their basic form or root. Stop word removal is another commonly employed technique, where insignificant terms such as articles or prepositions are eliminated to focus on meaningful words. For specific languages, tailored techniques are applied, such as diacritic removal for Arabic or the use of specialized dictionaries and linguistic databases to standardize terms. Additionally, advanced methods like NER are sometimes used to identify key elements such as proper nouns, dates, or locations, enriching data quality for more detailed analysis. These preprocessing steps are common to most studies. Fig. 10, given below, describes the most common preprocessing techniques used.

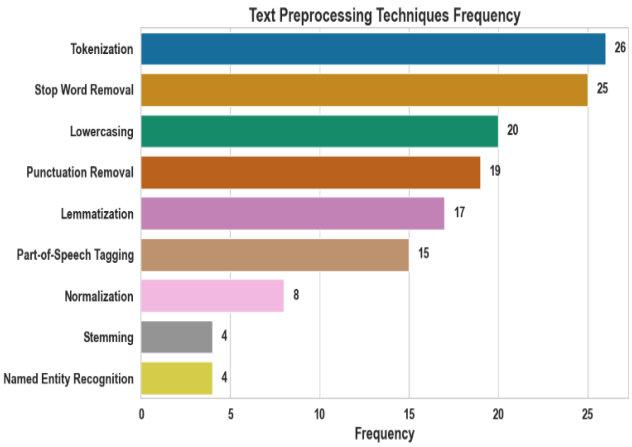


Fig. 10. Most commonly used preprocessing techniques.

2) Analysis of Language Pairs: This section provides an in-depth examination of each language pair, with a particular focus on the datasets employed, the applied techniques, and the performance achieved, while highlighting their unique features and the specific challenges they pose for processing and modeling. The analysis encompasses several key aspects of cross-language plagiarism detection, including the use of translation, the integration of multilingual semantic resources, the methods adopted for feature extraction, and the classification strategies implemented. It also covers the similarity measures applied, the datasets utilized, the levels of granularity considered (ranging from sentence-level to paragraph-level and document-level corpora), and the performance metrics reported. From a global perspective, CLPD research shows a reliance on a set of widely used datasets such as PAN-PC-11, SemEval, JRC-Acquis, Wikipedia, Europarl, and collections of conference papers, with their frequency of use often reflecting the language pairs investigated. In the following sections, we provide a more detailed distribution of datasets for English–Arabic and English–Spanish, which represent the most extensively studied language combinations.

a) English-Arabic language pair: For the En-Ar language pair, the preprocessing methods include tokenization, normalization, removal of stop words (RSW), and diacritic removal for Arabic text. Feature extraction techniques include pre-trained models like Word2Vec, Skip-Gram, CBOW, and TF-IDF. ML models, such as SVM, LR, DT, RF, KNN, and deep neural networks (DNN), are employed, with CS serving as the primary similarity measure. In most studies, Arabic texts are translated into English using automatic translation tools like Google Translate, followed by similarity computations using measures such as CS, ED, or semantic measures like WUP and LCH. For studies utilizing the SemEval-2017 dataset, which focuses on sentences, the combination of translation (Word2Vec+TF-IDF+MUSE) and the SVC model for classification yielded notable results with an F1-score of 87%. PC for the same combination also achieved a score of 81.47%. In datasets like OPUS, KSUCCA, and EAPCOUNT,

integrating BabelNet with a DNN architecture achieved a significant accuracy of 97.01%. Similarly, the use of translation, the fingerprint approach, and the Winnow algorithm with the LLR technique achieved an accuracy of 97%. Some studies also focused solely on measuring similarity percentages between two sentences.

Fig. 11 reveals that the most commonly used datasets for En-Ar language pair research are SemEval-2017 and EAPCOUNT, highlighting their importance for evaluating semantic similarity models and multilingual plagiarism detection. SemEval-2017 is sentence-based, making it suitable for fine-grained similarity tasks, while Wikipedia, being document-based, is more relevant for applications requiring diverse and larger contexts. Additionally, datasets such as articles, tweets, and academic works demonstrate their value in approaches that require varied and realistic text sources. OPUS and KSUCCA, though moderately used, remain significant for aligned multilingual corpora, whereas books and MultiUN complement the range by addressing specific needs. This distribution reflects a preference for datasets that balance accessibility, diversity, and specificity.

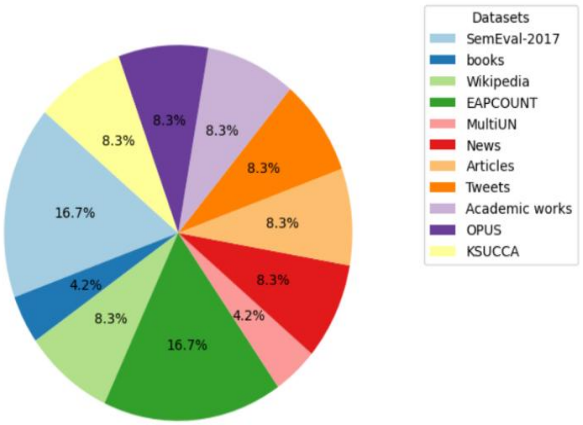


Fig. 11. Most used datasets for the English-Arabic language pair.

b) English-Spanish language pair: For the En-Es language pair, various methodologies, datasets, and performance results are highlighted. The most important approaches leverage advanced transformer models and deep learning architectures. For example, the use of mBART for feature extraction and SLSTM for classification achieved an accuracy of 97.94%, precision of 98.57%, recall of 97.47%, and an F1-score of 98.01%. This method is evaluated using datasets such as Europarl, PAN-PC-11, Wikipedia, JRC-Acquis, and Conference papers. Another notable method used several embedding techniques, including Doc2Vec, GloVe, FastText, BERT, Word2Vec, and Sen2Vec, combined with SLSTM. This strategy achieves accuracy rates between 98.41% and 99.81% on datasets such as JRC-Acquis, PAN-PC-11, Europarl, and Wikipedia. In contrast, methods using traditional feature extraction techniques like TF-IDF and Word2Vec also showed impressive performance. These methods realized plagiarism detection rates of 83.5% and 84.2% on PAN-PC-11 and PAN-PC-12, respectively, and a precision score of 44.3% on SemEval-2017, indicating their efficiency in simpler scenarios. However, approaches centered on knowledge graphs, such as those using Wikidata or BabelNet, displayed moderate results. For example, the method based on Wikidata achieved a plagiarism detection rate of 57.3%, with a precision of 72.3% and a recall of 47.4%. In comparison, the BabelNet-based method recorded a plagiarism detection rate of 60.87%, precision of 70.36%, and recall of 53.99%, both analyzed on the PAN-PC-11 dataset. Similarly, another approach that employed BiGRU and GNN in combination with WordNet and the CS measure achieved a plagiarism detection rate of 62%, precision of 20.3%, and recall of 8.5%.

Fig. 12 illustrates the distribution of datasets used for En-Es. The PAN-PC-11 dataset accounts for the largest proportion at 25%, indicating its extensive use in research. Other datasets, including PAN-PC-12, PAN-PC-14, SemEval-2016, and SemEval-2017, each represent 12.5%, reflecting their consistent contribution to CLPD tasks. Notably, Europarl is used at a smaller rate, 8.3%, while datasets like JRC-Acquis, Wikipedia, MRPC, and conference papers contribute 4.2% each.

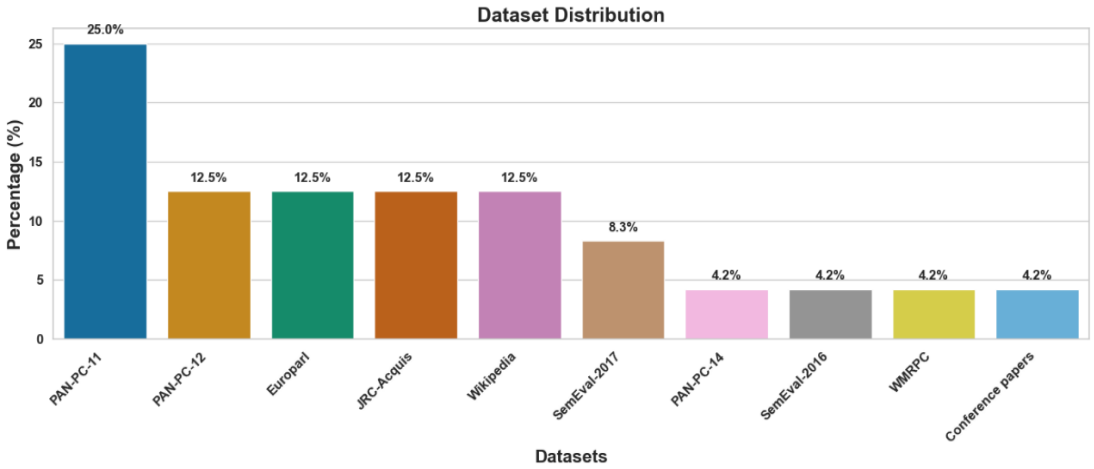


Fig. 12. Most used datasets for the English-Spanish language pair.

c) English-German language pair: For the En-De language pair, the approaches encompass both translation-based and translation-independent methods. The PAN-PC-11 dataset is the most frequently utilized, with the CS measure being widely applied across approaches. Among translation-independent methods, the use of mBART and SLSTM achieves an accuracy of 95.59%, a precision of 95.21%, a recall of 96.85%, and an F1-score of 96.02% across datasets such as PAN-PC-11, JRC-Acquis, Europarl, Wikipedia, and conference papers. For translation-based methods, XLM-R and M-BERT deliver notable results, achieving an accuracy of 87.2% and an F1-score of 90.8% on the GLUE dataset. Conversely, approaches leveraging knowledge graphs, such as Wikidata and BabelNet, demonstrate moderate performance. For instance, the Wikidata-based method achieves a plagiarism detection rate of 52.1%, precision of 67.2%, and recall of 42.5%, while the BabelNet-based approach records a plagiarism detection rate of 52.96%, precision of 63.06%, and recall of 46.71% on the PAN-PC-11 dataset. Additionally, translation-independent methods using traditional feature extraction techniques, including TF-IDF and Word2Vec, show strong performance, with plagiarism detection rates of 85.7% and 86.2% on the PAN-PC-11 and PAN-PC-12 datasets, respectively. In conclusion, transformer models, particularly mBART and SLSTM, demonstrate the most significant results for the En-De language pair, substantially outperforming traditional and knowledge-graph-based approaches.

d) English-French language pair: For the En-Fr language pair, the best-performing approach, translation-independent, combines the mBART transformer for feature extraction with the SLSTM model, using the CS measure. This method achieves exceptional results on datasets such as PAN-PC-11, JRC-Acquis, Europarl, Wikipedia, and conference papers, with an accuracy of 98.83%, a precision of 98.42%, a recall of 99.32%, and an F1-score of 98.87%. Another notable translation-independent approach utilizes CBOW for feature extraction, DT as classifiers, and the CS measure. Tested on datasets including Wikipedia, conference papers, product reviews, Europarl, and JRC-Acquis, this method achieves a solid F1-score of 89.15%. Additionally, a translation-independent approach based on the creation of a knowledge graph using WordNet, combined with BiGRU and GNN models and the CS measure, shows moderate performance. It achieves a precision of 50.6%, a recall of 69%, and a plagiarism detection rate of 58.4% when tested on datasets such as Europarl, JRC-Acquis, and Wikipedia. Another approach, using XLM-R and M-BERT on the GLUE dataset, attains an accuracy of 86.2% and an F1-score of 90.2%.

e) English-Vietnamese language pair: For the English-Vietnamese language pair, both translation-based and translation-independent strategies are emphasized. Translation-based models, such as XLM-R and M-BERT, utilize multilingual embeddings to achieve notable results on the GLUE dataset, with an accuracy of 84.3% and an F1 score of 88.5%. In contrast, translation-free methods leveraging Word2Vec embeddings and lexical resources like WordNet and VietNet demonstrate strong performance, with the SLSTM model achieving an accuracy of 89.61% on the TED dataset. Meanwhile, without word embedding techniques, the use of BabelNet and VietNet with deep learning models (SLSTM) achieves impressive results on the TED dataset, with a precision of 80.3%, a recall of 95.8%, and an F1 score of 87.4%. These findings underscore the effectiveness of both approaches, with translation-free methods often delivering comparable or superior performance.

f) English-Chinese language pair: For the English-Chinese language pair, the approach begins with the use of the Skip-Gram model of Word2Vec and ED as the similarity measure, which attained a high H-score of 97.09% on the NDLTD dataset without relying on translation. Meanwhile, translation-independent methods employing fingerprints and WordNet with the DC deliver a solid performance, achieving a precision of 87% and a recall of 78% on the CNKI dataset. On the other hand, translation-based methods, such as the combination of Google Translate and RCNN with Word2Vec, demonstrate an accuracy of 88.87% on the student dataset. Additionally, the use of Google Translate and BABYLON alongside TF-IDF yielded a Mean Average Precision (MAP) score of 500 on the Xinhua English news dataset.

g) English-Russian language pair: For the English-Russian language pair, all the approaches rely on transformer-based models for the feature extraction phase, achieving results above 80%. The approach using BERT achieves a precision of 96%, outperforming those based on LASER, SE, WS, NMT+1-Gram, and NMT+2-Gram. Furthermore, the use of FastText with CS on the Wikipedia dataset achieves a precision of 83%, a recall of 79%, and an F-score of 80%. Likewise, no use of MSN.

h) English-Persian language pair: For the EN-PE language pair, various translation tools and techniques are implemented, including English experts, Google Translate, and bilingual translators. One notable approach uses the Mizan dataset with WuP as a similarity measure, achieving an accuracy of 98.82%. Google Translate combined with BILBOWA and other techniques like T+MA, CL-LSI, and CL-ESA, leveraging BabelNet for feature extraction, delivers varying precision rates, ranging from 21% to 83% on the HAMTA-CL dataset. The bilingual translator approach employs VSM and TF-IDF with CS, achieving a precision of 88%, a recall of 96%, and an F-score of 91% on Internet sources. Additionally, transformer-based methods like XLM-R, M-BERT, and DistilBERT show interesting precision and similarity scores, with XLM-R achieving 95.62% and 95.17%, M-BERT achieving 91.88% and 91.55%, and DistilBERT achieving 89.51% and 89.08% on the PESTS dataset.

i) English-Urdu language pair: For the English-Urdu language pair, the first approach, based on Google Translate, WordNet, and ML models (KNN, SVM, DT, RF, NB), applied to CLPD-UE-19 with a document-level granularity, achieves an accuracy of 92% with KNN. In contrast, the n-grams + T+MA approach, tested on CLEU (sentence/document level), shows lower performance, with an F1-score of 73.2% for binary classification and 55.2% for ternary classification.

j) English-Indonesian language pair: For the English-Indonesian language pair, two translation-based methods are presented. The first method utilizes Google Translate combined with fingerprinting techniques and the Winnow algorithm, achieving an accuracy of 84.7% on a dataset of scientific articles using the Jaccard similarity measure. The second method relies on a dictionary database, combined with advanced techniques such as LSA, LVQ, and SVD, along with WordNet and the CS measure. This second approach outperforms the first, achieving an accuracy of 87% on the same dataset.

k) Underexplored language Pairs: Certain languages receive limited attention in existing research, primarily because they are not widely spoken or used in international communication. The analysis highlights the underexplored language pairs in CLPD systems, showcasing efforts to address these pairs despite limited resources and research focus. English-Japanese (EN-JA) and English-Korean (EN-KO) achieved promising accuracies of 87.49% and 87.26%, respectively, though with limited datasets. For Traditional Chinese (Ti-Ch), translation-independent methods (Doc2Vec, SLSTM) showed lower precision (54-55%) on datasets like CWMT and SemEval2014. French-Arabic (Fr-Ar), using WordNet similarity measures on datasets of news articles and academic works, recorded a WUP score of 63%. These results reflect the challenges and opportunities for advancing CLPD for less-resourced language pairs.

V. Discussion

This section presents the approaches employed in CLPD, provides a summary of our SLR, and highlights promising research directions that remain largely unexplored.

A. General Architecture for CLPD

In this SLR, we examine the approaches used for CLPD, categorized into four main groups. Similarity-based approaches (MSN) utilize lexical-semantic resources such as WordNet and BabelNet to enhance semantic comparison and measure the relationship between words across languages. These methods also rely on similarity metrics such as cosine similarity, Jaccard index, and Wu-Palmer similarity to assess textual resemblance. Traditional models like TF-IDF and Word2Vec offer a statistical representation of texts. Fingerprinting techniques create distinctive signatures to detect similarities. Lastly, Transformer and deep learning frameworks, including BERT, XLM-R, and mBART, utilize advanced neural architectures for more efficient plagiarism detection.

Fig. 13 illustrates the steps of CLPD systems, which are based on two main strategies: with translation and without translation. In the first approach, the text is translated into a common language to facilitate direct comparison, while in the second approach, detection is performed using multilingual representations without translation. After this initial stage, the text is represented using three methods: Word Embedding (WE alone or WE + MSN), where models like Word2Vec, BERT, or XLM-R transform the text into vectors, either autonomously or by integrating the Multilingual Semantic Network (MSN) to enhance multilingual understanding; Fingerprinting (FP alone or FP + MSN), which generates unique textual fingerprints that can be used alone or combined with MSN to improve semantic alignment; and MSN alone, used independently to detect similarity between texts. After establishing the vector representation of the text, two outputs are obtained: the first output gives a similarity score of two texts via a similarity measure, while the second is a classification using ML or DL algorithms, employing models such as SVM, KNN, or neural networks to distinguish plagiarized from non-plagiarized texts. Finally, a model evaluation phase is carried out using performance measures such as accuracy, precision, recall, and F1 score to assess the effectiveness of the proposed model.

B. Results and Future Research Avenues

This part outlines the key results of the SLR for each inquiry, emphasizing the conclusions derived from the examined studies. It also suggests potential avenues for future investigation, intend to address the identified gaps and steer further advancements in the area.

RQ1: Based on the analysis of the literature review, we have identified 20 language pairs used in work on CLPD. The most frequently studied pairs are En-Ar, En-Es, and En-De, respectively, as shown in Fig. 3. Although other combinations were mentioned, some, such as the English-Hindi pair, were excluded from our selection because the corresponding studies did not meet the inclusion criteria (non-open-source research, review, or monolingual plagiarism). Thus, the total number of pairs actually exploitable remains limited to ten. English appears to be the dominant pivot language, being present in 98% of the language combinations identified.

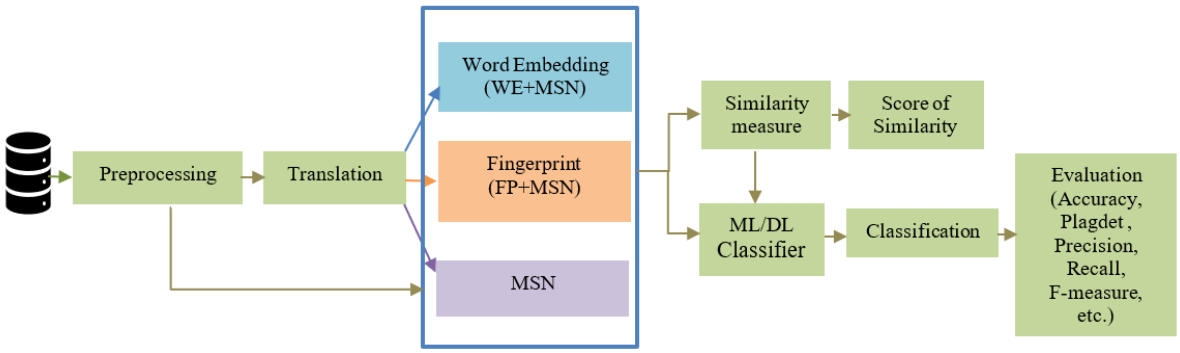


Fig. 13. General architecture for CLPD systems.

RQ2: Concerning the use of translation for CLPD, we observe that during the timeframe of 2014-2025, most documents are translated into the same language, with English being the predominant language for these translations. Additionally, there is a notable portion of studies that are directly incorporated without the process of translation. Among the most widely used translation tools, Google Translate dominates, as illustrated in Fig. 4 and Fig. 5. We also note that for the least explored language pairs, translation is the most used, compared to the most tiled language pair in the context of the CLPD system. According to our analysis, studies that have adopted embedding techniques without going through translation also give significant results, particularly those that have used transformers for different language pairs. The mBART model produced interesting results exceeding 95% for the three language pairs En-Es, En-Fr, and En-De, while the Bert model achieved 96% accuracy for En-Ru. In addition, studies using translation tools achieved interesting results for the English-Persian language pair, with the use of an English expert for translation achieving 98.82% accuracy. In general, the use of advanced embedding techniques, such as transformers, yields interesting results in many languages without the need for translation.

RQ3: Through our review of the literature, we have pinpointed four primary categories of multilingual representation techniques employed in CLPD systems: 1) conventional methods (such as TF-IDF, Word2Vec, FastText), 2) transformer and deep learning architectures (such as mBERT, mBART, XLM-R), 3) multilingual semantic networks like WordNet, BabelNet, VietNet, and Wikidata, and 4) fingerprinting. Historically, conventional methods have been the most prevalent, whereas fingerprinting represents the least investigated area, as depicted in Fig. 6. An analysis of trends over time (Fig. 7) indicates a remarkable transition towards Transformer-based models beginning in 2018, leading to a decline in the use of traditional approaches due to their superior performance and capacity to manage intricate multilingual scenarios. Transformer-based models significantly exceed the performance of traditional word embedding methods, which are often constrained in their F1-measure, as illustrated in Table IV. Furthermore, these models have been evaluated using an expanding range of language pairs, highlighting their versatility. Regarding multilingual semantic networks, WordNet is the most commonly utilized, followed closely by BabelNet, VietNet, and Wikidata (Fig. 8). Nevertheless, these methodologies remain underutilized in CLPD. Lastly, fingerprints are the least frequently employed.

RQ4: Analysis of existing work reveals a diversity of similarity metrics used in CLPD. Nevertheless, the majority of studies favor cosine similarity, due to its effectiveness with vector representations and its ability to evaluate similarities regardless of text length. The second most widely used metric is the Jaccard index, mainly applied to the comparison of sets of words or characters. Other measures, such as Wu-Palmer (WuP), LCS (Longest Common Subsequence), ED (Euclidean Distance), Lin similarity, or the Dice coefficient, appear more marginally, often in specific contexts or as complements to the main methods. Finally, the Containment measure is the least represented, as shown in Fig. 9.

RQ5: In our study, we analyzed several features for each language pair used in CLPD, namely: translation strategy, MSN, feature extraction techniques, methods used (ML/DL), similarity measures, datasets exploited, data granularity, as well as performance obtained. Existing work in CLPD generally pursues two main objectives: similarity score calculation, aimed at comparing sentences or documents using different metrics to evaluate feature extraction techniques, and automatic classification, which consists of training machine or deep learning models to detect cases of plagiarism, relying on metrics such as precision, recall, or F1-measure. Several factors significantly influence the performance of these approaches. Firstly, the granularity of the data plays a decisive role: an analysis at the document level is often more effective than a global analysis of the sentence. Secondly, two main strategies are adopted for multilingual processing: translation into a pivot language, usually English, prior to vectorization, or direct use of the original texts via multilingual models such as mBERT or XLM-R. The choice of method (multilingual semantic networks, fingerprints, traditional embeddings, or Transformers-type models) also has a major impact on the quality of results. Furthermore, poorly endowed languages pose a real challenge due to the lack of suitable linguistic resources, which limits model performance. Finally, advanced models such as BERT, SLSTM, mBART, or XLM-R stand out for their ability to learn rich contextual representations, offering better performance than traditional methods.

Nonetheless, CLPD approaches are still facing several unresolved challenges that manifest across various dimensions of current research:

- Construct diversified multilingual datasets, especially for under-represented languages, and study new and less explored language pairs.
- Analyze and compare the outcomes of methods that incorporate translation against those that do not, while assessing the reliability and performance of various machine translation tools.
- Assess model robustness to paraphrasing, focusing on their capacity to recognize semantically equivalent expressions despite significant syntactic or lexical alterations.
- Examine multiple MSNs such as WordNet, BabelNet, and Wikidata, comparing their performance in terms of semantic richness, contextual precision, and multilingual support.
- Investigate the integration of multiple MSNs (e.g., combining WordNet with BabelNet or Wikidata) to enhance lexical coverage, especially in the context of low-resource languages or domain-specific terminology.
- Evaluate other similarity measures, including LCS, Wu-Palmer, Euclidean distance, Lin, Dice, and cosine similarity, using various types of textual representations.

- Analyze the influence of textual granularity (sentence-level, paragraph-level, or document-level) on detection accuracy and consistency.
- Enhance the capabilities of multilingual knowledge graph-based approaches, focusing on improving their adaptability, scalability, and semantic precision across languages.
- Expand the use of multilingual transformers, given their underutilization in current CLPD studies, to improve scalability and detection accuracy.

VI. Conclusion

Detecting multilingual plagiarism is a major challenge, especially in the academic field. This issue needs a deep study to identify the approaches used for the CLPD system and to respond to many questions. In this context, our work presents an SLR for multilingual plagiarism detection covering the period from 2014 to 2025. Our study proposes four types of multilingual text representations: traditional approaches, multilingual semantic networks, fingerprinting methods, and deep learning models. This review highlights several key findings. The English-Arabic language pair emerges as the most frequently studied, and English appears in 98% of the examined language pairs. Over 60% of the studies incorporate a translation phase, with Google Translate being the most commonly used tool. The mBART model has shown promising results, achieving over 95% accuracy for the En-Es, En-Fr, and En-De language pairs, while the BERT model reached 96% accuracy for the En–Ru pair. Additionally, studies involving translation tools report strong performance for the En-Pe language pair, with an English expert tool for translation reaching up to 98.82% accuracy. Overall, the use of advanced embedding techniques such as transformers has yielded strong results across various language pairs, often without requiring translation. However, it is important to note that the best results do not depend only on the approach used. Firstly, the granularity of the dataset plays a crucial role: detecting plagiarism in a document is different from detecting plagiarism in a sentence. Secondly, the choice of language pair has a significant impact: little-studied languages generally achieve poorer results due to a lack of linguistic resources and research attention. Thirdly, the type of the adopted approach influences the performance. All these factors have a significant impact on CLPD system performance. For future work, we aim to develop universal CLPD models capable of handling any language and form of plagiarism. A promising direction is to leverage knowledge graph–based text representations, enabling the detection of subtle paraphrasing and improving cross-lingual generalization.

References

[1] T. Foltýnek, N. Meuschke, et B. Gipp, “Academic Plagiarism Detection: A Systematic Literature Review”, ACM Comput. Surv., vol. 52, no 6, p. 1-42, nov. 2020, doi: 10.1145/3345317.

[2] O. Zimba et A. Y. Gasparyan, “Plagiarism detection and prevention: a primer for researchers”, doi: 10.5114/reum.2021.105974.

[3] H. E. Mostafa et F. Benabbou, “A deep learning based technique for plagiarism detection: a comparative study”, IAES Int. J. Artif. Intell. IJ-AI, vol. 9, no 1, Art. no 1, mars 2020, doi: 10.11591/ijai.v9.i1.pp81-90.

[4] S. Sierra-Martínez, M.-E. Martínez-Figueira, M. D. Castro Pais, et T. Pessoa, “You work, I copy’. Images, narratives and metaphors around academic plagiarism through Fotovoz”, Br. Educ. Res. J., vol. 50, no 3, p. 1514-1532, 2024, doi: 10.1002/berj.3977.

[5] S. P. J. M. (Serge) Horbach et W. (Willem) Halffman, “The extent and causes of academic text recycling or ‘self-plagiarism’”, Res. Policy, vol. 48, no 2, p. 492-502, mars 2019, doi: 10.1016/j.respol.2017.09.004.

[6] F. Alvi, M. Stevenson, et P. Clough, “Paraphrase type identification for plagiarism detection using contexts and word embeddings”, Int. J. Educ. Technol. High. Educ., vol. 18, no 1, p. 42, août 2021, doi: 10.1186/s41239-021-00277-8.

[7] K. Avetisyan, G. Gritsay, et A. Grabovoy, “Cross-Lingual Plagiarism Detection: Two Are Better Than One”, Program. Comput. Softw., vol. 49, no 4, p. 346-354, août 2023, doi: 10.1134/S0361768823040138.

[8] O. Bakhteev et al., “Cross-Language Plagiarism Detection: A Case Study of European Languages Academic Works”, in Academic Integrity: Broadening Practices, Technologies, and the Role of Students: Proceedings from the European Conference on Academic Integrity and Plagiarism 2021, S. Bjelobaba, T. Foltýnek, I. Glendinning, V. Krásničan, et D. H. Dlabolová, Éd., Cham: Springer International Publishing, 2022, p. 143-161. doi: 10.1007/978-3-031-16976-2_9.

[9] M. AlSallal, R. Iqbal, S. Amin, A. James, et V. Palade, “An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection”, in 2016 9th International Conference on Developments in eSystems Engineering (DeSE), août 2016, p. 203-208. doi: 10.1109/DeSE.2016.1.

[10] M. Singh et V. Gupta, “Review of Extrinsic Plagiarism Detection Techniques and Their Efficiency Comparison”, in Advanced Network Technologies and Intelligent Computing, I. Woungang, S. K. Dhurandher, K. K. Pattanaik, A. Verma, et P. Verma, Éd., Cham: Springer International Publishing, 2022, p. 609-624. doi: 10.1007/978-3-030-96040-7_46.

[11] M. AlSallal, R. Iqbal, V. Palade, S. Amin, et V. Chang, “An integrated approach for intrinsic plagiarism detection”, Future Gener. Comput. Syst., vol. 96, p. 700-712, juill. 2019, doi: 10.1016/j.future.2017.11.023.

[12] M. F. Manzoor, M. S. Farooq, M. Haseeb, U. Farooq, S. Khalid, et A. Abid, “Exploring the Landscape of Intrinsic Plagiarism Detection: Benchmarks, Techniques, Evolution, and Challenges”, IEEE Access, vol. 11, p. 140519-140545, 2023, doi: 10.1109/ACCESS.2023.3338855.

[13] R. Pandit, S. Sengupta, S. K. Naskar, N. S. Dash, et M. M. Sardar, “Improving Semantic Similarity with Cross-Lingual Resources: A Study in Bangla—A Low Resourced Language”, Informatics, vol. 6, no 2, Art. no 2, juin 2019, doi: 10.3390/informatics6020019.

[14] D. G. Dusza, “Machine Translation in the Writing Process: Pedagogy, Plagiarism, Policy, and Procedures”, in Second Handbook of Academic Integrity, S. E. Eaton, Éd., Cham: Springer Nature Switzerland, 2024, p. 1487-1509. doi: 10.1007/978-3-031-54144-5_152.

[15] V. Vysotska, Y. Burov, V. Lytvyn, et A. Demchuk, “Defining Author’s Style for Plagiarism Detection in Academic Environment”, in 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), août 2018, p. 128-133. doi: 10.1109/DSMP.2018.8478574.

[16] J. P. Wahle, T. Ruas, T. Foltýnek, N. Meuschke, et B. Gipp, “Identifying Machine-Paraphrased Plagiarism”, in Information for a Better World: Shaping the Global Future, M. Smits, Éd., Cham: Springer International Publishing, 2022, p. 393-413. doi: 10.1007/978-3-030-96957-8_34.

[17] H. Ezzikouri, M. Erritali, et M. Oukessou, “Plagiarism Detection in Across Less Related Languages (English-Arabic): A Comparative Study”, in Smart Data and Computational Intelligence, F. Khoukhi, M. Bahaj, et M. Ezziyyani, Éd., Cham: Springer International Publishing, 2019, p. 207-213. doi: 10.1007/978-3-030-11914-0_22.

[18] F. M. Prentice et C. E. Kinden, “Paraphrasing tools, language translation tools and plagiarism: an exploratory study”, Int. J. Educ. Integr., vol. 14, no 1, p. 11, déc. 2018, doi: 10.1007/s40979-018-0036-7.

[19] Nirbhay Kumar Chaubey Et Al, “Automatic plagiarism detection and extraction in a multilingual: a critical study and comparison”, janv. 2022, doi: 10.17605/OSF.IO/DWUK4.

[20] M. U. Akhtar, J. Liu, Z. Xie, X. Liu, S. Ahmed, et B. Huang, “Entity alignment based on relational semantics augmentation for multilingual knowledge graphs”, Knowl.-Based Syst., vol. 252, p. 109494, sept. 2022, doi: 10.1016/j.knosys.2022.109494.

[21] M. Elkhidir, M. M. Ibrahim, T. A. Khalid, S. Ibrahim, et M. Awadalla, “Plagiarism detection using free-text fingerprint analysis”, in 2015 World

Symposium on Computer Networks and Information Security (WSCNIS), sept. 2015, p. 1-4. doi: 10.1109/WSCNIS.2015.7368306.

[22] M. A. Botto Tobar, M. G. J. van den Brand, et A. Serebrenik, "Cross-Language Plagiarism Detection: Methods, Tools, and Challenges: A Systematic Review", Int. J. Adv. Sci. Eng. Inf. Technol., vol. 12, no 2, p. 589-599, mai 2022, doi: 10.18517/ijaseit.12.2.14711.

[23] N. Alotaibi et M. Joy, "Using Sentence Embedding for Cross-Language Plagiarism Detection", in Artificial Intelligence XXXVII, M. Bramer et R. Ellis, Éd., Cham: Springer International Publishing, 2020, p. 373-379. doi: 10.1007/978-3-030-63799-6_28.

[24] H. Aljuaid, "Cross-Language Plagiarism Detection using Word Embedding and Inverse Document Frequency (IDF)", Int. J. Adv. Comput. Sci. Appl. IJACSA, vol. 11, no 2, Art. no 2, 32/29 2020, doi: 10.14569/IJACSA.2020.0110231.

[25] J. Ferrero, F. Agnes, L. Besacier, et D. Schwab, "CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity", 5 avril 2017, arXiv: arXiv:1704.01346. doi: 10.48550/arXiv.1704.01346.

[26] C.-M. Chang, C.-H. Chang, et 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, doi: 10.1002/pra2.229.

[27] R. Kuznetsova, "CrossLang: the system of cross-lingual plagiarism detection", janv. 2019.

[28] N. Alotaibi et M. Joy, "English-Arabic cross-language plagiarism detection", in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online: INCOMA Ltd, sept. 2021, p. 44-52. doi: 10.26615/978-954-452-072-4_006.

[29] J. Ferrero, F. Agnes, L. Besacier, et D. Schwab, "UsingWord Embedding for Cross-Language Plagiarism Detection", 10 février 2017, arXiv: arXiv:1702.03082. doi: 10.48550/arXiv.1702.03082.

[30] R. Lachraf, E. M. B. Nagoudi, Y. Ayachi, A. Abdelali, et D. Schwab, "ArbEngVec : Arabic-English Cross-Lingual Word Embedding Model", in Proceedings of the Fourth Arabic Natural Language Processing Workshop, W. El-Hajj, L. H. Belguith, F. Bougares, W. Magdy, I. Zitouni, N. Tomeh, M. El-Haj, et W. Zaghouani, Éd., Florence, Italy: Association for Computational Linguistics, août 2019, p. 40-48. doi: 10.18653/v1/W19-4605.

[31] M. Roostaee, S. M. Fakhrahmad, et M. H. Sadreddini, "Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection", Expert Syst. Appl., vol. 160, p. 113718, déc. 2020, doi: 10.1016/j.eswa.2020.113718.

[32] N. Ehsan et A. Shakery, "Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information", Inf. Process. Manag., vol. 52, no 6, p. 1004-1017, nov. 2016, doi: 10.1016/j.ipm.2016.04.006.

[33] H. Y. Chen et P. Vines, "Multi Queries Methods of the Chinese-English Bilingual Plagiarism Detection", Appl. Mech. Mater., vol. 462-463, p. 1158-1162, 2014, doi: 10.4028/www.scientific.net/AMM.462-463.1158.

[34] N. Ehsan, A. Shakery, et F. Tompa, "Cross-lingual text alignment for fine-grained plagiarism detection", J. Inf. Sci., vol. 45, p. 016555151878769, août 2018, doi: 10.1177/0165551518787696.

[35] M. Roostaee, M. H. Sadreddini, et 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, mars 2020, doi: 10.1016/j.ipm.2019.102150.

[36] H. Asghari, O. Fatemi, S. Mohtaj, H. Faili, et P. Rosso, "On the use of word embedding for cross language plagiarism detection", Intell. Data Anal., vol. 23, no 3, p. 661-680, janv. 2019, doi: 10.3233/IDA-183985.

[37] S. Enayati, "Introducing an Automated Technique for Bilingual Plagiarism detection of English-Persian Documents", 2014.

[38] L. T. Nguyen et D. Dien, "Vietnamese- English Cross-Lingual Paraphrase Identification Using Siamese Recurrent Architectures", in 2019 19th International Symposium on Communications and Information Technologies (ISCIT), sept. 2019, p. 70-75. doi: 10.1109/ISCIT.2019.8905116.

[39] A. A. P. Ratna et al., "Cross-Language Plagiarism Detection System Using Latent Semantic Analysis and Learning Vector Quantization", Algorithms, vol. 10, p. 69, juin 2017, doi: 10.3390/a10020069.

[40] E. Hattab, "Cross-Language Plagiarism Detection Method: Arabic vs. English", in 2015 International Conference on Developments of E-Systems Engineering (DeSE), déc. 2015, p. 141-144. doi: 10.1109/DeSE.2015.25.

[41] J. Min, "Cross-Language Translation Algorithm Based on Word Vector and Syntactic Analysis", Int. J. Multiphysics, vol. 18, no 2, Art. no 2, avr. 2024.

[42] I. Muneer, M. Sharjeel, M. Iqbal, R. M. A. Nawab, et P. Rayson, "CLEU - A Cross-language english-urdu corpus and benchmark for text reuse experiments", J. Assoc. Inf. Sci. Technol., vol. 70, no 7, p. 729-741, 2019, doi: 10.1002/asi.24074.

[43] M. AlMousa, R. Benlamri, et R. Khoury, "A novel word sense disambiguation approach using WordNet knowledge graph", Comput. Speech Lang., vol. 74, p. 101337, juill. 2022, doi: 10.1016/j.csl.2021.101337.

[44] R. Navigli et S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network", Artif. Intell., vol. 193, p. 217-250, déc. 2012, doi: 10.1016/j.artint.2012.07.001.

[45] H. Ezzikouri, M. Oukessou, M. Youness, et 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, in ICCBDC '18. New York, NY, USA: Association for Computing Machinery, août 2018, p. 22-27. doi: 10.1145/3264560.3264562.

[46] M. M. Zahid, K. Abid, A. Rehman, M. Fuzail, et N. Aslam, "An Efficient Machine Learning Approach for Plagiarism Detection in Text Documents", J. Comput. Biomed. Inform., vol. 4, no 02, Art. no 02, mars 2023, doi: 10.56979/402/2023.

[47] J. Stegmüller, F. Bauer-Marquart, N. Meuschke, T. Ruas, M. Schubotz, et B. Gipp, "Detecting Cross-Language Plagiarism using Open Knowledge Graphs", p. 853881 Bytes, 2021, doi: 10.6084/m9.figshare.17212340.v3.

[48] F. Safi-Esfahani, S. Rakian, et M. H. Nadimi-Shahraki, "English-Persian Plagiarism Detection based on a Semantic Approach", J. AI Data Min., vol. 5, no 2, p. 275-284, juill. 2017, doi: 10.22044/jadm.2016.770.

[49] M. Franco Salvador, P. Gupta, et P. Rosso, "Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing", 2014, p. 227-236. doi: 10.1007/978-3-642-54798-0_12.

[50] M. H. A. Almayali, Z. Alaa, et S. Tiun, Cross Language Plagiarism of Arabic-English Documents Using Linear Log. Noor Publishing, 2017.

[51] H. Ezzikouri, M. Oukessou, M. Erritali, et Y. Madani, "Fuzzy Cross Language Plagiarism Detection Approach Based on Semantic Similarity and Hadoop MapReduce", in Recent Advances in Intuitionistic Fuzzy Logic Systems: Theoretical Aspects and Applications, S. Melliani et O. Castillo, Éd., Cham: Springer International Publishing, 2019, p. 181-190. doi: 10.1007/978-3-030-02155-9_15.

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

[53] M. Al-Suhaiqi, M. A. S. Hazaa, et M. Albared, "Arabic English Cross-Lingual Plagiarism Detection Based on Keyphrases Extraction, Monolingual and Machine Learning Approach", Asian J. Res. Comput. Sci., vol. 2, no 3, Art. no 3, févr. 2019, doi: 10.9734/ajrcos/2018/v2i330075.

[54] L. Gang, Z. Quan, et L. Guang, "Cross-language plagiarism detection based on WordNet", in Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence, in ICIAI '18. New York, NY, USA: Association for Computing Machinery, mars 2018, p. 163-168. doi: 10.1145/3194206.3194222.

[55] M. Mentari, I. Rozi, et M. Rahayu, "Cross-Language Text Document Plagiarism Detection System Using Winnowing Method", J. Appl. Intell. Syst., vol. 7, p. 44-57, mai 2022, doi: 10.33633/jais.v7i1.5950.

[56] A. Aljohani et M. Mohd, “Arabic-English Cross-language Plagiarism Detection using Winnowing Algorithm”, Inf. Technol. J., vol. 13, p. 2349-2355, déc. 2014, doi: 10.3923/itj.2014.2349.2355.

[57] H. V. T. Chi, D. L. Anh, N. L. Thanh, et D. Dinh, “English-Vietnamese Cross-Lingual Paraphrase Identification Using MT-DNN”, Eng. Technol. Appl. Sci. Res., vol. 11, no 5, Art. no 5, oct. 2021, doi: 10.48084/etasr.4300.

[58] M. Alshehri, N. Beloff, et M. White, “AraXLM: New XLM-RoBERTa Based Method for Plagiarism Detection in Arabic Text”, in Intelligent Computing, K. Arai, Éd., Cham: Springer Nature Switzerland, 2024, p. 81-96. doi: 10.1007/978-3-031-62277-9_6.

[59] C. Bouaine et F. Benabbou, “Efficient cross-lingual plagiarism detection using bidirectional and auto-regressive transformers”, IAES Int. J. Artif. Intell. IJ-AI, vol. 13, no 4, Art. no 4, déc. 2024, doi: 10.11591/ijai.v13.i4.pp4619-4629.

[60] V. ZubarevD et V. SochenkovI, “Cross-language text alignment for plagiarism detection based on contextual and context-free models”, 2019.

[61] M. Abdous, P. Piroozfar, et B. MinaeiBidgoli, “PESTS: Persian_English cross lingual corpus for semantic textual similarity”, Lang. Resour. Eval., août 2024, doi: 10.1007/s10579-024-09759-3.

[62] W. Bao, J. Dong, Y. Xu, Y. Yang, et X. Qi, “Exploring Attentive Siamese LSTM for Low-Resource Text Plagiarism Detection”, Data Intell., vol. 6, no 2, p. 488-503, mai 2024, doi: 10.1162/dint_a_00242.

[63] O. Hourrane et E. H. Benlahmar, “Graph transformer for cross-lingual plagiarism detection”, IAES Int. J. Artif. Intell. IJ-AI, vol. 11, no 3, Art. no 3, sept. 2022, doi: 10.11591/ijai.v11.i3.pp905-915.

[64] C. Bouaine, F. Benabbou, et I. Sadgali, “Word Embedding for High Performance Cross-Language Plagiarism Detection Techniques”, Int. J. Interact. Mob. Technol. IJIM, vol. 17, no 10, Art. no 10, mai 2023, doi: 10.3991/ijim.v17i10.38891.

[65] D. Dinh et N. L. Thành, “English–Vietnamese cross-language paraphrase identification using hybrid feature classes”, J. Heuristics, vol. 28, avr. 2022, doi: 10.1007/s10732-019-09411-2.

[66] T. Ter-Hovhannisyan et K. Avetisyan, “Transformer-Based Multilingual Language Models in Cross-Lingual Plagiarism Detection”, in 2022 Ivannikov Memorial Workshop (IVMEM), sept. 2022, p. 72-80. doi: 10.1109/IVMEM57067.2022.9983968.

[67] I. Muneer, N. Waheed, M. A. Ashraf, et R. M. Adeel Nawab, “Sentential Cross-lingual Paraphrase Detection for English-Urdu Language Pair”, Eur. J. Artif. Intell., p. 30504554251319446, mars 2025, doi: 10.1177/30504554251319446.

[68] J. Memon, M. Sami, R. A. Khan, et M. Uddin, “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)”, IEEE Access, vol. 8, p. 142642-142668, 2020, doi: 10.1109/ACCESS.2020.3012542.

[69] L. E. Jiani, S. E. Filali, E. H. B. Lahmar, et I. Haloum, “Deep Learning-Based Approaches Using Medical Imaging for Therapy Response Prediction in Breast Cancer: A Systematic Literature Review”, Int. J. Online Biomed. Eng. IJOE, vol. 20, no 12, Art. no 12, sept. 2024, doi: 10.3991/ijoe.v20i12.49709.

[70] L. E. Jiani, S. E. Filali, E. H. B. Lahmar, et I. Haloum, “Deep Learning-Based Approaches Using Medical Imaging for Therapy Response Prediction in Breast Cancer: A Systematic Literature Review”, Int. J. Online Biomed. Eng. IJOE, vol. 20, no 12, Art. no 12, sept. 2024, doi: 10.3991/ijoe.v20i12.49709.