IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 4, December 2024, pp. 4619~4629
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i4.pp4619-4629

Efficient cross-lingual plagiarism detection using bidirectional and auto-regressive transformers

Chaimaa Bouaine, Faouzia Benabbou
Department of Mathematics and Computer Science, Faculty of Science Ben M'Sick, Hassan II University, Casablanca, Morocco

Article Info

Article history:

Received Nov 28, 2023
Revised Jan 29, 2024
Accepted Mar 15, 2024

Keywords:

Cross-language plagiarism
English-French
English-German
English-Spanish
Multilingual bidirectional and auto-regressive transformers
Siamese long short-term memory

ABSTRACT

The pervasive availability of vast online information has fundamentally altered our approach to acquiring knowledge. Nevertheless, this wealth of data has also presented significant challenges to academic integrity, notably in the realm of cross-lingual plagiarism. This type of plagiarism involves the unauthorized copying, translation, ideas, or works from one language into others without proper citation. This research introduces a methodology for identifying multilingual plagiarism, utilizing a pre-trained multilingual bidirectional and auto-regressive transformers (mBART) model for document feature extraction. Additionally, a siamese long short-term memory (SLSTM) model is employed for classifying pairs of documents as either "plagiarized" or "non-plagiarized". Our approach exhibits notable performance across various languages, including English (En), Spanish (Es), German (De), and French (Fr). Notably, experiments focusing on the En-Fr language pair yielded exceptional results, with an accuracy of 98.83%, precision of 98.42%, recall of 99.32%, and F-score of 98.87%. For En-Es, the model achieved an accuracy of 97.94%, precision of 98.57%, recall of 97.47%, and an F-score of 98.01%. In the case of En-De, the model demonstrated an accuracy of 95.59%, precision of 95.21%, recall of 96.85%, and F-score of 96.02%. These outcomes underscore the effectiveness of combining the MBART transformer and SLSTM models for cross-lingual plagiarism detection.

Corresponding Author:

Chaimaa Bouaine
Department of mathematics and Computer Science, Faculty of Science Ben M'Sick, Hassan II University
N°51, Hay Ifriquia Rue Echahid Eloualid Essaghir, Casablanca 20000, Morocco
Email: chaimaa.bouaine-etu@etu.univh2c.ma

1. INTRODUCTION

In today's interconnected world, where the exchange of ideas and information knows no geographical boundaries, multilingual plagiarism detection has become a big challenge in the academic and intellectual landscape. Unlike traditional plagiarism, which generally involves the unauthorized use of content in a single language, multilingual plagiarism transcends linguistic boundaries. It involves the appropriation, translation, or reproduction of an original work in one language and its unauthorized use in another, raising important ethical and practical questions [1]. Plagiarism takes different forms, each characterized by nuances of severity and intent. Copy-paste occurs when text is reproduced verbatim from a source without proper attribution. Plagiarism of ideas occurs when concepts or ideas are borrowed without proper credit. Self-plagiarism occurs when one's own previous work is submitted without proper citation or authorization. Paraphrasing, the rewriting of another's work without proper attribution or using overly similar words and sentence structures, is also a form of plagiarism [2]. When exploring the field of cross-language plagiarism, it is imperative to understand that it goes far beyond the translation of documents from one language to another. The cross-language plagiarism encompasses a variety of plagiarism techniques, including direct plagiarism, paraphrasing, semantic and even plagiarism involving a change in style. In a multilingual context, the complexity of plagiarism is accentuated by linguistic and cultural differences between languages, adding an extra dimension to the challenge of plagiarism detection. Translating a document from one language to another can lead to significant changes in style and grammatical structure.

This phenomenon calls for heightened awareness, innovative detection methods, and a better understanding of the subtleties of plagiarism as it cross linguistic (CL) and cultural boundaries. In response to this complex issue, plagiarism detection technology has undergone significant progress, characterized by the integration of advanced machine learning [3] and deep learning (DL) models [4], [5], with a strong focus on transformers. Transformers are a DL architecture, that relies on the parallel multi-head attention mechanism. They have been used successfully in various natural language processing (NLP) tasks thanks to their ability to capture long-range dependencies in data. These models excel in understanding the context and nuanced meaning of words within a sequence, making them suitable for identifying cases of multilingual plagiarism [6]. The transformer architecture comprises two parts: the encoder and the decoder, the encoder captures the essence of the input text, generating a representation that takes into account language-specific features and structures. At the same time, the decoder is responsible for producing an output sequence, using the encoded information from the input to ensure that language-specific details and structures are preserved in the output [7].

Concurrently, long short-term memory (LSTM) networks signify an evolution in recurrent neural networks, specializing in capturing long-term temporal dependencies. In the context of plagiarism detection, LSTMs are instrumental, in transforming integration vectors of each text into dense representations that capture semantic and structural nuances. These representations undergo combination through a fusion layer, and the similarity between texts is assessed using cosine distance. The use of LSTMs facilitates effective modeling of intricate relationships between words and phrases, thereby enhancing the detection of similarities and contributing significantly to plagiarism detection across diverse linguistic contexts [8]. LSTM is employed for various types of plagiarism, including paraphrase plagiarism [9] and plagiarism of ideas [10]. The multilingual bidirectional and auto-regressive transformers (mBART) model have been used for different NLP tasks [11], [12], but not for cross-language plagiarism.

In this research paper, we introduce an innovative approach for cross-language plagiarism detection (CLPD) based mBART. Additionally, we employ the siamese long short-term memory (SLSTM) model. This combination of mBART and SLSTM proves highly effective in addressing the challenge of CLPD and offers a robust solution for detecting and understanding plagiarism across language barriers.

The remaining sections of this document are structured as follows: the subsequent section outlines the current state of the art in CLPD techniques employing transformer models. Section 3 provides a detailed description of our proposed methodology. The experimentation and corresponding results are expounded upon in section 4. Ultimately, we conclude by summarizing the main findings and exploring potential directions for future research.

2. RELATED WORK

Various methodologies have been developed to detect instances of cross-language plagiarism. This form of plagiarism encompasses various manipulations employed to plagiarize a text, including paraphrasing, style alteration, idea replication, or straightforward translation. Given their success in diverse areas of NLP and machine learning, we are particularly interested in approaches based on the transformer architecture, such as bidirectional encoder representations from transformers (BERT), as well as embedding models like GloVe, and Word2vec. Our analysis includes essential pre-processing steps that are pivotal in enhancing the accuracy of plagiarism detection. Additionally, we explore similarity measures between two documents to further refine our assessment. Furthermore, since only a limited number of studies have employed transformers for multilingual plagiarism, we extend this literature review to encompass plagiarism within the same language.

Chi *et al.* [13] proposed a CLPD system specifically designed for English-Vietnamese (En-Vi) to identify paraphrases. Their approach utilized the multi-task deep neural network (MTDNN) model, incorporating pre-trained models such as multilingual bag-of-words (BOW), enhanced sequential inference model (ESIM), and multilingual bidirectional encoder representations from transformers (M-BERT) along with XLM-RoBERTa (XLM-R). Through fine-tuning on the GLUE datasets, XLM-R demonstrated superior performance compared to M-BERT, achieving an accuracy of 84.3% and an F-score of 88.5%. Abdous *et al.* [14] tackled the issue case of Persian-English CL plagiarism. They employed multilingual transformer models including XLM-R, M-BERT, and DistilBERT, and computed the semantic similarity between Persian and English text with the cosine similarity. The findings are that the XLM-RoBERTa model achieved a Pearson correlation (PC) of 95.62% on the PESTS dataset compared to M-BERT which achieved a correlation of 91.88%, while DistilBERT achieved a correlation of 89.51%. Avetisyan *et al.* [15] proposed a multilingual plagiarism detection method based on candidate retrieval, text preprocessing, the use of inverted indexing, and the evaluation of similarity between texts to determine whether they are translations of each other. To speed up the search, an inverted index is used, and candidates are ranked according to their relevance using the Okapi BM25 ranking function. In addition, a word frequency-based approach was developed to create multilingual concept clusters for plagiarism detection using universal WordNet (UWN) as a corpus. Using BERT to assess the similarity between texts and determine whether they are translations of each other. The fine-tuning process for XLM-RoBERTa involved using a dataset generated from English Wikipedia and scientific papers covering diverse topics sourced from Google Scholar. The approach achieved respectively an F-score of 81%, 82% for English-Russian, English-Spanish, and 81% for English-French, English-German, and English-Hungarian (En-Hu) using the vMRPC dataset.

Further research included text alignment method by Zubarev *et al.* [16] presented a CL plagiarism for Russian-English languages within the context of plagiarism detection. This method includes neural machine translation (NMT) to translate Russian text into English. A comparative analysis of various models, including sentence embedding, BERT, word substitution (WS), and LASER, is conducted to identify translated plagiarism. BERT model, achieved a precision rate of 96%, recall of 93%, and an F-score of 95%. They then evaluated the similarity between these translated sentences using Jaccard metrics, both with 1-grams (NMT) and 2-grams (NMT2). The NMT model yielded a precision rate of 85%, recall of 80%, and a F-score of 82. Additionally, logistic regression (LR) was employed in two configurations: LR-1, where all techniques were utilized, and LR-2, where sentence embedding and word substitution techniques were applied. The LR-1 achieved a precision rate of 91%, a recall of 80%, and an F-score of 85%. Yang *et al.* [17] tackled the challenge of paraphrase identification using three distinct models including BOW, ESIM, and BERT. The experimental results underscore the significance of BERT, with an accuracy of 93.8%, 90.8%, 90.7%, 89.2%, 85.4%, 83.1%, and 83.9% respectively for English, French, Spanish, German, Chinese, Japanese, and Korean on PAWS-X dataset.

In another study, Hattab [18] tackled the issue of English-Arabic CL plagiarism using the concept of latent semantic indexing (LSI). LSI is used to construct a CL semantic space to assess the contextual similarity between two documents from English and Arabic. The LSI method achieved a 93% similarity rate when applied to the English-Arabic parallel corpus of united nations texts (EAPCOUNT), but the authors did not provide the performance of the method in usual metrics such as accuracy or precision. Zahid *et al.* [19] introduced a method for identifying plagiarism in documents written in Urdu-English languages. They assessed similarity using Jaccard, Cosine, and longest common subsequence (LCS) with both n-grams and trigrams. Five machine learning classifiers, including k-nearest neighbors (KNN), naïve bayes (NB), support vector machines (SVM), decision trees, and random forest were trained on the CLPD-UE-19 dataset to construct predictive models. KNN classifier produced outcomes, achieving an accuracy of 93% for training, 92% for testing, and 84% using cross-validation.

Gupta *et al.* [20] proposed a cross linguistic conceptual thesaurus similarity (CL-CTS) method to measure semantic similarity between concepts in language pairs such as English-Spanish and English-German. This approach yielded recall above 90% for Spanish-English and above 80% for English-German on three datasets: JRC-Acquis, PAN-PC-11, and Wikipedia. Wahle *et al.* [21] addressed the problem of paraphrase plagiarism detection, in order to differentiate between text generated by humans and text generated by machine paraphrasing tools like SpinBot and SpinnerChief. Various embedding techniques were investigated such as classical ones including GloVe, Word2vec, Doc2vec, and FastText and eight techniques based on the Transformer architecture including BERT, RoBERTa, a lite BERT (ALBERT), distillable BERT (DistilBERT), efficiently learning an encoder that classifies token replacements accurately (ELECTRA), BART, eXtreme learning network (XLNet), and longformer for identifying machine-paraphrases. Three machine learning classifiers, namely SVM, NB, and LR, were examined, whereas GloVe+SVM produces noteworthy results across all datasets (arXiv, theses, Wikipedia). The technique, Longformer, achieved an average F1-Micro of 99.7% for SpinBot and F1-Micro of 71.6% for SpinnerChief, while human evaluators achieved F1-Micro 78.4% for SpinBot and an F-score of 65.6% for SpinnerChief.

This state-of-the-art shows that the interest in transformer-based techniques is growing, and that detection performance achieved using different transformers is very promising. Table 1 provides a comprehensive overview of the transformer models, summarizing their performance on the basis of four key characteristics. It ranks the models according to the type of plagiarism including (CL, monolingue (ML)), the language pairs studied (Russian-English (Ru-En), Persian-English (Pe-En)) and their best performance measures, including F1-Micro (F1-M) and PC. The dataset column describes the datasets used, and the final column describes the references (Refs) for each model.

The Table 1 summarizes the best performance for all state-of-the-art models based on transformers. Several language pairs were processed to detect the CL cases. For the pre-processing phase, the studies adopted the basic pre-processing techniques such as stemming, stop word removal, tokenization, lower-casing, and punctuation removal.

Table 1. Transformer models performance

Transformers	Plagiarism	Language pairs	Best performance	Dataset	Refs.
M-BERT	CL	Pe-En	PC: 91.88%	PESTS	[14]
XLM-RoBERTa	CL	Pe-En	PC: 95.62%	PESTS	[14]
BERT	CL	Ru-En	P: 96%		[14]
	Paraphrasing/ML	English	F1-M: 99.44	arXiv_Papers	[21]
RoBERTa	Paraphrasing/ML	English	F1-M: 99.05%	arXiv_Papers	[21]
ALBERT	Paraphrasing/ML	English	F1-M: 98.91	arXiv_Papers	[21]
DistilBERT	Paraphrasing/ML	English	F1-M: 99.32	arXiv_Papers	[21]
BART	Paraphrasing/ML	English	F1-M: 99.58	arXiv_Papers	[21]
ELECTRA	Paraphrasing/ML	English	F1-M: 99.20	arXiv_Papers	[21]
XLNet	Paraphrasing/ML	English	F1-M: 99.65	arXiv_Papers	[21]
Longformer	Paraphrasing/ML	English	F1-M: 99.38	arXiv_Papers	[21]

It is obvious that the transformers' predominant application lies in monolingual plagiarism detection, with remarkable results given that all models achieved a proficiency level of over 90%. This performance not only underlines the transformers' outstanding capabilities, but also places them as superior to word embedding models such as GloVe, Word2Vec, FastText, and Doc2Vec when coupled with classifiers such as SVM, NB, and LR. In particular XLM-RoBERTa, demonstrated exceptional performance compared to M-BERT in different language pairs. For example, for the En-Vi pair, XLM-RoBERTa obtained an F-score of 88.5%, and for the Persian-English pair, a PC of 95.62% was obtained. BERT also achieved 96% accuracy for the Russian-English pair. These results underline the effectiveness of transformer-based models in different language pairs. As for the BART model, it is mainly used for monolingual plagiarism and also provided significant results in the paraphrasing task. Cosine similarity is extensively utilized in the majority of studies for computing the similarity between documents, although alternative measures like the Dice coefficient and Jaccard similarity have also found application.

3. METHOD

Expanding upon the baseline, our proposal involves employing mBART in an innovative approach for CLPD focusing on three language pairs: En-Es, En-Fr, and En-De. The methodology comprises several steps, including data preprocessing, feature extraction utilizing the mBART transformer, and classification based on the SLSTM model. To evaluate the performance of the approach, we apply it to five distinct datasets, namely PAN-11, JRC-ACQUIS, Wikipedia, EUROPARL, and conference papers, using them for both model training and performance assessment.

As illustrated in Figure 1, our model incorporates two distinct input layers, input layer_1 and input layer_2, designed to receive sequences from different languages within the training data. The second layer employs a shared embedding layer based on the mBART model, representing input sequences with an output dimension of 1024. Two branches of the model process the two respective input sequences, and their representations are then directed to a shared LSTM layer, generating an 80-dimensional output. The Siamese neural network architecture is tailored for discerning similarity between two multilingual text sequences. The outputs of these LSTMs are flattened to yield one-dimensional vectors. These vectors undergo various transformations and combinations, including subtraction, multiplication, and cosine distance operations, with the goal of capturing diverse aspects of similarity between the two sequences for classification as either plagiarized (P) or not plagiarized (NP). Ultimately, the extracted features are concatenated and processed by dense layers with rectified linear unit (ReLU) activations, culminating in an output layer using a sigmoid activation function to predict binary similarity.

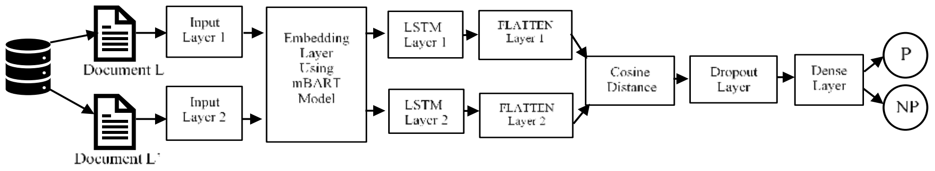


Figure 1. Proposed methodology

3.1. Dataset

In this paper, we used five datasets PAN11 [22], JRC-ACQUIS [23], EUROPARL [24], Wikipedia [25], and conference papers [26] for three language pairs En-Fr, En-Es, En-De. In the case of English-French, we used 10,620 plagiarized (P) documents, combining data from two sources: conference papers and JRC-Acquis. As our problem is binary classification, we include in our dataset 9,442 pairs of non-plagiarized documents taken from the Europarl dataset, as well as documents taken from Wikipedia. For English-Spanish, the dataset considered includes 10,210 plagiarized documents, resulting from the merging of the two datasets Pan-11 and JRC-Acquis, and 9,423 non-plagiarized documents from the Europarl dataset and Wikipedia. In the case of English-German, we collected 12,221 plagiarized documents by combining data from the Pan-11 and JRC-Acquis datasets and we added 9,999 documents from the Europarl dataset and Wikipedia. This procedure enabled us to create complete and balanced datasets for each language pair. Table 2 summarizes the characteristics of each dataset used.

Table 2. Features of dataset used

Language pairs	DATASET	Pairs of documents	Target	Number total of documents
English-Frensh	Conference papers	620	P	20062
	JRC-Acquis	10000	P	
	Wikipedia+Europarl	9442	NP	
English-German	JRC-Acquis	9999	P	22220
	PAN-11	2222	P	
	Wikipedia+Europarl	9999	NP	
English-Spanish	PAN-11	210	P	19633
	JRC-Acquis	10000	P	
	Wikipedia+Europarl	9423	NP	

3.2. Preprocessing

The pre-processing step improves data quality by making it more consistent, understandable, and ready to use in the plagiarism detection process. It also facilitates the extraction of significant features and contributes to more accurate and efficient analysis to improve model performance. In our approach, we remove punctuation and eliminate characters such as commas and question marks. After that, the tokenization is applied to documents in order to obtain individual words or "tokens" for more granular analysis. We then proceed to remove empty words for each language, thus excluding commonly used words that generally have no significant meaning in the context of the document. Text normalization was conducted by converting text entirely to lowercase, ensuring that words written in different cases were treated consistently. Finally, decontraction techniques are employed to handle language-specific contractions, transforming, for example, "can't" into "can not" in English or "no es" into "no es" in Spanish. This ensures consistency in word representation and facilitates further text processing. In addition, to support the lemmatization of documents in ES-EN, FR-EN, and En-DE the natural language toolkit (NLTK) library in Python is utilized. This comprehensive pre-processing approach enables us to prepare the data in a consistent way.

3.3. Feature extraction

This step is performed by using an mBART transformer to produce vector representations, often referred to as embeddings or encodings of input data. These vector representations capture the semantic and contextual information present in input data and are used to facilitate various NLP tasks. Transformers showed a high capacity for encoding complex contextual and semantic information, making them well suited to many NLP tasks such as NMT [27], sentiment analysis [28], and bots' detection [29]. By passing data through the layers of a transformer model, it learns to automatically extract relevant features and create high-dimensional embeddings that retain valuable information for downstream tasks such as NLP, including plagiarism detection [30]. mBART a multilingual variant of the BART model, conceived to support many languages. It is a sequence-to-sequence language model structured with an encoder-decoder architecture. It undergoes pre-training on diverse data sources, including Wikipedia and others, across multiple languages, encompassing a total of 50 languages. The model's strength lies in its capability to capture intricate semantic representations and generalize across diverse linguistic contexts. Recognized for its effectiveness, mBART has become a preferred choice in various NLP applications. Widely employed in tasks such as automatic translation, automatic summarization, text generation, and question answering, mBART consistently demonstrates its prowess in applications requiring a sophisticated understanding of multilingual linguistic nuances [31].

3.4. Siamese long short-term memory

The LSTM serves as a specialized adaptation of recurrent neural networks, specifically designed to address the challenge of gradient vanishing. This characteristic makes LSTMs particularly effective in tasks requiring long-term memory, such as predicting text sequences. A Siamese network is a type of NN architecture that involves two input fields used to compare two patterns, and it produces a single output representing the similarity between these patterns. It operates by utilizing two distinct sub-networks to process each input pattern independently, thereby extracting relevant features. Subsequently, the cosine of the angle between the two resulting feature vectors is computed, serving as a measure of their similarity and, effectively, a distance value [32]. In our methodology, we harness the SLSTM model to improve the learning of plagiarism instances, enhancing the precision of document representations and adeptly identifying similarities between pairs of objects, particularly excelling in tasks related to similarity.

3.5. Hyperparameters of the proposed model

The hyperparameters of the model play a pivotal role in elevating its performance. We employed the mBART model with the identifier "facebook/mbart-large-50-many-to-many-mmt." This model generates an output vector of length 1024 and is engineered to handle input data of 250054 dimensions. This configuration underscores the model's complexity and its adeptness at efficiently processing a diverse array of linguistic data and NLP tasks. Regarding the dropout layer, we determined that a 40% dropout rate was optimal. This rate struck a balance, ensuring accuracy while preventing overfitting. The choice of the activation function is task-dependent. Given our binary classification problem with classes "plagiarized" (1) or "not plagiarized" (0), we opted for the binary cross-entropy loss function and the sigmoid activation function, which demonstrated efficiency for this particular problem. In terms of the optimizer, we conducted experiments with various options, including Adam, stochastic gradient descent (SGD), and Adadelta. Ultimately, Adadelta emerged as the most effective choice. To identify the optimal number of epochs and address overfitting concerns, we implemented early stopping from Keras. The hyperparameters for SLSTM+mBART are detailed in Table 3.

Table 3. Characteristics of SLSTM+mBART hyperparameters

Model	Hyperparameter	Value
mBART+SLSTM	Output Dim	1024
	Neuron	80
	dropout	0.4
	Activation function	Sigmoid
	Optimizer	Adadelta
	Loss function	Binary cross entropy
	Batch size	64
	Epochs	100

4. RESULTS AND DISCUSSION

In this section, we provide the results obtained using mBART and SLSTM for three language pairs: En-Es, En-Fr, and En-De. To assess performance comprehensively, our evaluation incorporates a series of key measures. These measures include accuracy (A), precision (P), recall (R), F1-score (F1), and area under the receiver operating characteristic curve (AUC).

4.1. Performance measure

The evaluation of the proposed models involves the utilization of performance metrics, which are derived by calculating true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) values. These metrics are crucial in assessing the model performance, and their calculation is presented as shown in (1) to (4).

$$A = (TN + TP)/(TN + FP + TP + FN) \tag{1}$$

$$P = TP/(TP + FP) \tag{2}$$

$$R = TP/(TP + FN) \tag{3}$$

$$F1 = 2 * (P * R)/(P + R) \tag{4}$$

The AUC is utilized in binary classification problems. It gauges how well a model can differentiate between positive and negative classes. In addition, the confusion matrix helps to identify areas of strength and those that may need improvement by using them to assess the precision of the models and their capacity to produce accurate predictions.

4.2. Results and analysis

In this section, we present our results using the mBART and SLSTM models for three language pairs: En-Es, En-Fr, and En-De. The Table 4 shows the performance detection rates of the approach for each language pair. To ensure rigorous evaluation, the dataset has been divided into three distinct parts: the training set with 60% of the data, the validation phase with 20% of the data, and finally, the test set, representing 20% of the data.

The results of the experiment, as presented in Table 4, demonstrate the model's commendable performance across all language pairs, with a notable emphasis on the En-Fr pair, showcasing an accuracy of 98.83%, precision of 98.42%, recall of 99.32%, F1 score of 98.87%, AUC of 99.84%, and an exceptionally low loss rate of 0.05. For the En-Es pair, the approach proves effective, attaining an accuracy of 97.94%, precision of 98.57%, recall of 97.47%, F1 score of 98.01%, AUC of 99.85%, and a loss rate of 0.07. In the case of the En-De pair, performance, though slightly lower, remains high, with a consistent accuracy of 95.59%, precision of 95.21%, recall of 96.85%, F1 score of 96.02%, AUC of 98.97%, and a loss rate of 0.15. Across all metrics and for the three language pairs, our approach yields successful outcomes. Metrics such as precision, recall, and F1 score exhibit high and consistent values between training, validation, and test sets. The AUC, reaching up to 98.97%, stands out as a robust indicator of performance, accompanied by the smallest observed loss of 0.05. In summary, the model exhibits commendable proficiency in plagiarism detection across all three language pairs, with outstanding performance observed for the En-Fr pair.

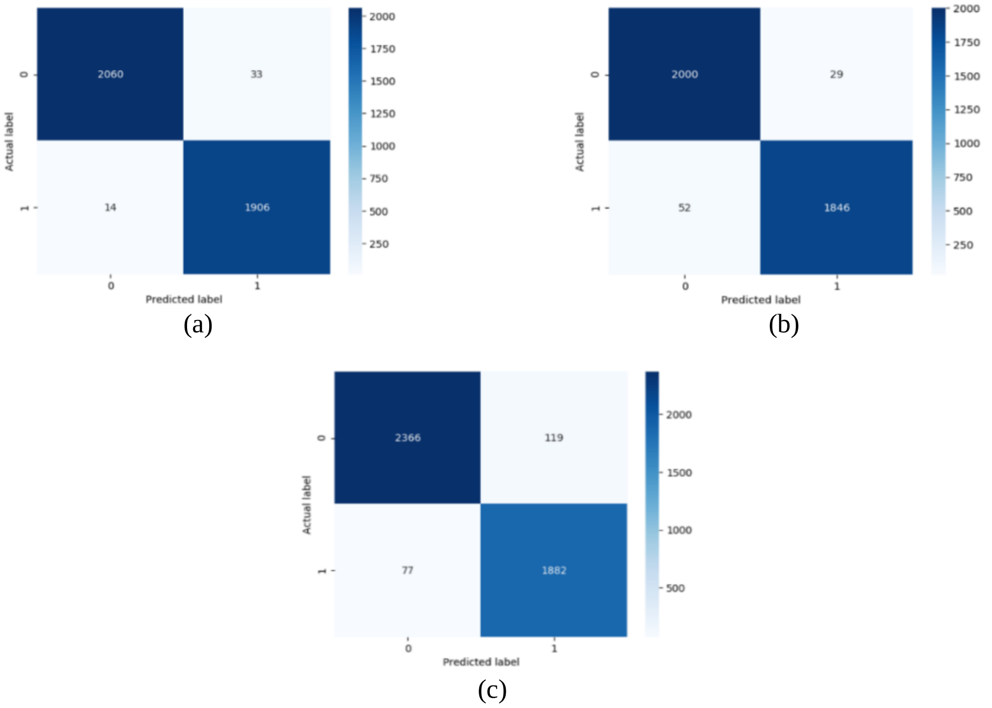
Table 4. Performances of our method for each language pair

Language Pairs	Metrics	Train (%)	Validation (%)	Test (%)
EN-ES	Accuracy	97.80	97.64	97.94
	Precision	98.25	98.24	98.57
	Recall	97.51	97.13	97.47
	F-score	97.87	97.68	98.01
	AUC	99.74	99.73	99.85
	Loss	0.08	0.08	0.07
EN-FR	Accuracy	98.68	98.57	98.83
	Precision	98.39	98.00	98.42
	Recall	99.15	99.36	99.32
	F-score	98.76	98.67	98.87
	AUC	99.82	99.79	99.84
	Loss	0.061	0.062	0.05
EN-DE	Accuracy	95.77	95.78	95.59
	Precision	94.99	95.05	95.21
	Recall	97.46	97.31	96.85
	F-score	96.21	96.16	96.02
	AUC	98.77	98.99	98.97
	Loss	0.16	0.15	0.15

Furthermore, Figure 2 provides the confusion matrices for our proposed approach, offering a detailed analysis of its performance across each language pair. In Figure 2(a), for the English-French language pair, the model correctly classifies 4013 documents, with only 33 instances classified as FN. Additionally, there are 14 cases of FP, indicating instances where the model failed to identify plagiarism. In Figure 2(b), for the English-Spanish language pair, the confusion matrix illustrates that out of a total of 3927 evaluated cases, 1846 were correctly identified as TP, representing plagiarism cases. Moreover, the model accurately recognized 2000 cases as TN. However, the model also produced 29 FN, signifying cases incorrectly classified as plagiarism, and 52 FP, indicating instances of actual plagiarism that the model failed to detect. In Figure 2(c), for the En-De language pair, the confusion matrix reflects a total of 4444 evaluated cases. The model demonstrated accurate recognition of 1882 plagiarism cases TP and 2366 non-plagiarism cases TN, showcasing commendable performance. Nonetheless, the model also made errors by classifying 119 cases as non-plagiarized when they were, in fact, plagiarized FN and identifying 77 cases as plagiarized when they were not FP.

To confirm the efficiency of our approach and make sure there is no overfitting, we examined the curves for accuracy and loss throughout the training as well as the validation phases which provide a close monitoring of the model's learning progress during these phases. In addition, they provide valuable information on the model's progress and enable the identification of potential problems such as over-fitting or under-fitting. The following figures show the accuracy and loss curves during the training, as well as the validation phases of our approach, based on mBART and SLSLTM models, applied to each language pair En-Fr, En-Es, and En-De.

Figure 3 shows that accuracy of both the training, as well as the validation phases for the En-Fr language pair is nearing perfection, achieving 98.68% in the case of the training set and 98.57% in the case of the validation set. Furthermore, loss curve demonstrates that the model converges effectively, reaching a minimum value of 0.061 in the case of training and 0.062 in the case of validation. The results indicate the model's strong predictive capabilities, with only a slight disparity between both training as well as validation data.



(a)

(b)

(c)

Figure 2. Confusion matrix for the mBART+SLSTM model for each language pair as (a) English-French language pair, (b) English-Spanish language pair, and (c) English-German language pair

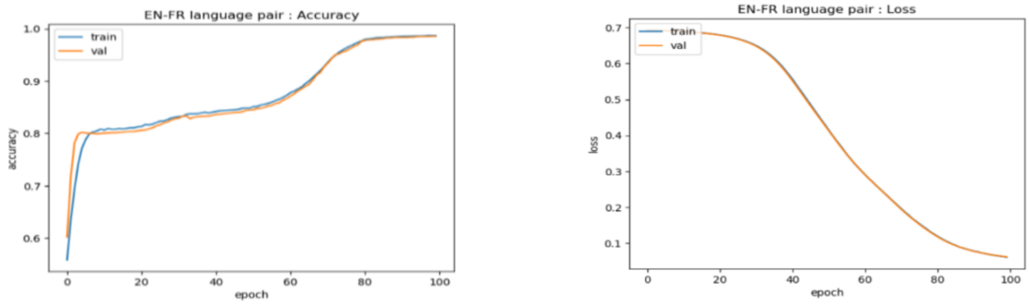


Figure 3. Evaluating accuracy and loss metrics in English-French language with mbart+SLSTM model

Figure 4 shows a good evolution of the accuracy and loss curve for the En-De language pair. The accuracy curve for the two phases of training as well as validation develops slowly in some epochs, with 95.77% for data in training and 95.78% for data in validation. The loss curve was minimized to 0.16 in training and 0.15 in validation, with minimal deviation, underlining the model's excellent performance.

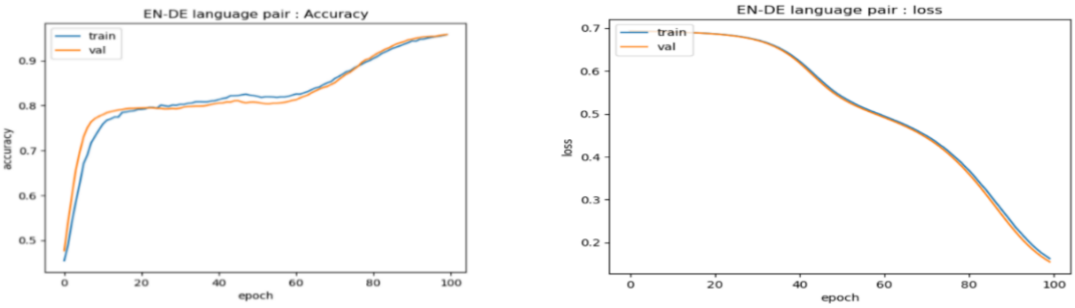


Figure 4. Evaluating accuracy and loss metrics in English-German language with mbart+SLSTM model

Figure 5 illustrates the progression of the model over 100 epochs for the En-Es language pair. The accuracy curve shows a rapid progression in the early epochs, followed by a more gradual evolution up to epoch 80, and finally, a rapid development in the later stages up to epoch 100, reaching impressive values of 97.80 in training and 97.64% in validation. The loss curve was minimized to 0.088 in training and 0.082 in validation.

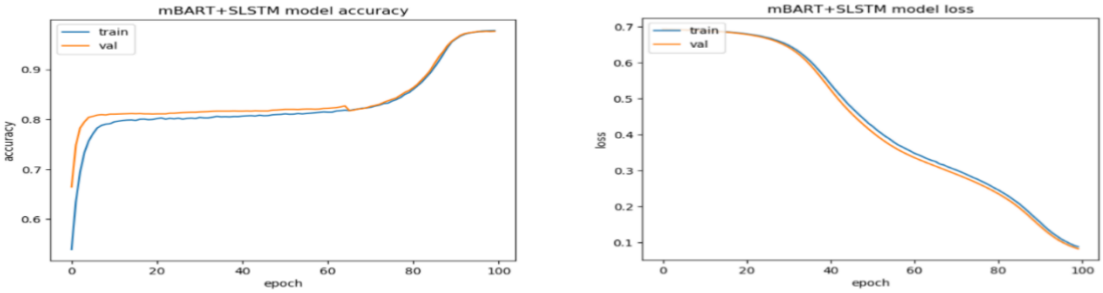


Figure 5. Evaluating accuracy and loss metrics in English-Spanish language with mbart+SLSTM Model

Finally, Table 5 gives a comparison of the performance achieved by the various transformers used in the literature on different datasets. Our model achieved outstanding results, surpassing those obtained by state-of-the-art models in the context of crosslingual plagiarism. While the BART model is very effective in detecting monolingual plagiarism [21], the multilingual BART model combined with the SLSTM model also outperformed in crosslingual plagiarism detection in three language pairs: (En-Fr), (En-Es), and (En-De). It is crucial to emphasize that this comparison serves as a reference, considering the models were not trained on the same dataset. It provides insight on models genralization to unknown data, assess their relative performance on CL plagiarism task, and help select the most appropriate transformer model for a given language pair. Our approach shows a good result with a precision of 98.83%, 97.94%, and 95.59% for En-Fr, En-Es, and En-De pairs respectively, while the combination of Bert and XLM-RoBERTa achieved only 81%. We should also mention that the large size and balancing of the training data played an important role in reaching these satisfactory performances.

Table 5. Transformer models comparison

Refs.	Transformers	Language pairs	Performance (%)	Dataset
[13]	XLM-RoBARTa	En-Vi	A :84.3, F1: 88.5	GLUE
[14]		Pe-En	PC: 95.62	PESTS
[13]	M-BERT	En-Vi	A:73.7, F1:81.3	GLUE
[14]		Pe-En	PC: 91.88%	PESTS
	DistilBert		PC: 89.51	
[14]	BERT	Ru-En	P: 96%	Negative-1
[15]	Bert+XLM-RoBERTa	En-Ru, En-Fr	P:72, F1:81	vMRPC
		En-Hu	P:74, F1:81	
		En-Es	P:72, F1:82	
		En-De	P:70, F:81	
Proposed approach	mBART+SLSTM	En-Fr	A: 98.83, F1: 98.87	PAN11, JRC-ACQUIS EUROPARL, WIKIPEDIA, Conference papers
		En-Es	A: 97.94, F1: 98.01	
		En-De	A: 95.59, F1: 96.02	

In summary, it is essential to emphasize that the nature of each language significantly influences the assessment of the performance of the plagiarism detection model. Each language has its own linguistic characteristics, grammatical structure, vocabulary, and cultural peculiarities, which influence the way plagiarism is performed and detected. Even so, the proposed method was capable of capturing the contexts of each of the languages handled. As part of this process, we took into account the linguistic nuances inherent in each language studied (English, French, Spanish, German). This led to the adjustment of our plagiarism detection model to make it sensitive to subtle variations in language use. This adaptation considerably enhances our model's ability to identify cross language plagiarism. The integration of the mBART transformer model has significantly improved the ability of our model to capture the contextual nuances of a given text. By capturing both syntactic and semantic context, it provides a comprehensive understanding that proves invaluable invaluable in discerning various forms of plagiarism. This improvement extends to the study of direct, semantic and paraphrastic instances, making the models more robust and versatile in the realm of multilingual plagiarism detection. The BART model excels in capturing the semantic nuances of documents due to its bidirectional architecture and auto-regressive capability. By analyzing text sequences in both directions, it manages to grasp the profound meaning of sentences, thus capturing semantic similarities between documents. The results confirm that our approach outperforms other works using different metrics. A limitation inherent in our approach lies in the restriction to just three language pairs, which could restrict the generalizability of our results to other language combinations. The complexity of the model we have used is another limitation, requiring a remarkable amount of time, especially in view of the substantial amount of data.

5. CONCLUSION

This study addresses the challenge of crosslingual plagiarism detection for three language pairs En-Es, En-Fr, and En-De. The proposition is based on the transformer architecture mBART and SLSTM model. The models were extensively trained on a combination of five publicly accessible corpora: Pan11, JRC-Acquis, Europarl, Wikipedia, and conference papers. The combination of the mBART transformer and SLSTM technique significantly enhanced the performance of plagiarism detection. Notably, our method outperformed in detecting plagiarism for the En-Fr language pair, achieving an accuracy of 98.83%. In addition, the combination of mBart for feature extraction and SLSTM neural network for learning the plagiarism cases provides a good accuracy of 95.59% for the En-De pair, and 97.94% for the En-Es pair during the test phase. The proposed approach is proving to be effective in comprehending sequential data and effectively maintaining extended connections among words. In our upcoming research, our objective is to analyze other language pairs such as Arabic-English and Arabic-Frensh. Another goal is to reduce the complexity of these models, as they currently require significant processing time, and we intend to explore other contextual integration models based on the transformer architecture to achieve more efficiency and faster execution.

ACKNOWLEDGEMENTS

We extend our heartfelt thanks to the research team: cloud computing, network and systems engineering for their invaluable support and encouragement during our research endeavors. Furthermore, we express gratitude to Hassan II University for granting us the opportunity to explore the expansive realm of artificial intelligence through our studies.

REFERENCES

[1] Z. Ceska, M. Toman, and K. Jezek, “Multilingual plagiarism detection,” in *Artificial Intelligence: Methodology, Systems, and Applications*, vol. 5253, Springer: Berlin, Heidelberg, pp. 83–92, 2008, doi: 10.1007/978-3-540-85776-1_8.

[2] S. Awasthi, “Plagiarism and academic misconduct: a systematic review,” *DESIDOC Journal of Library and Information Technology*, vol. 39, no. 2, pp. 94–100, 2019, doi: 10.14429/djlit.39.2.13622.

[3] M. Maqbool, I. Hanif, S. Iqbal, and A. Shabbir, “Optimized feature extraction and cross-lingual text reuse detection using ensemble machine learning models,” *Research Square*, 2022, doi: 10.21203/rs.3.rs-2122778/v1.

[4] S. Alzahrani and H. Aljuaid, “Identifying cross-lingual plagiarism using rich semantic features and deep neural networks: a study on Arabic-English plagiarism cases,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1110–1123, 2022, doi: 10.1016/j.jksuci.2020.04.009.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Advances in Neural Information Processing Systems*, vol. 4, pp. 3104–3112, 2014.

[6] O. Hourrane and E. H. Benlahmar, “Graph transformer for cross-lingual plagiarism detection,” *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 905–915, 2022, doi: 10.11591/ijai.v11.i3.pp905-915.

[7] A. Vaswani *et al.*, “Attention is all you need,” *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[9] H. Shahmohammadi, M. H. Dezfoulian, and M. Mansoorizadeh, “Paraphrase detection using LSTM networks and handcrafted features,” *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 6479–6492, 2021, doi: 10.1007/s11042-020-09996-y.

[10] F. Benabbou, H. E. Mostafa, and E. M. Hambi, “A system for ideas plagiarism detection: state of art and proposed approach,” *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, pp. 81–90, 2020, doi: 10.11591/ijai.v9.i1.

[11] A. C. Stickland, X. Li, and M. Ghazvininejad, “Recipes for adapting pre-trained monolingual and multilingual models to machine translation,” *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 3440–3453, 2021, doi: 10.18653/v1/2021.eacl-main.301.

[12] J. G. M. FitzGerald, “STIL -- simultaneous slot filling, translation, intent classification, and language identification: initial results using mBART on MultiATIS++,” in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 576-581, 2020.

[13] H. V. T. Chi, D. L. Anh, N. L. Thanh, and D. Dinh, “English-Vietnamese cross-lingual paraphrase identification using MT-DNN,” *Engineering, Technology and Applied Science Research*, vol. 11, no. 5, pp. 7598–7604, 2021, doi: 10.48084/etasr.4300.

[14] M. Abdous, P. Piroozfar, and B. M. Bidgoli, “PESTS: persian_english cross lingual corpus for semantic textual similarity,” *Language Resources and Evaluation*, 2024, doi: 10.1007/s10579-024-09759-3.

[15] K. Avetisyan, A. Malajyan, T. Ghukasyan, and A. Avetisyan, “A simple and effective method of cross-lingual plagiarism detection,” *Research Square*, 2023, doi: 10.21203/rs.3.rs-3040948/v1.

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

[17] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge, “PAWS-X: a cross-lingual adversarial dataset for paraphrase identification,” *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing,* Association for Computational Linguistics, pp. 3687–3692, 2019, doi: 10.18653/v1/d19-1382.

[18] E. Hattab, “Cross-language plagiarism detection method: Arabic vs. English,” *2015 International Conference on Developments in eSystems Engineering, DeSE 2015*. IEEE, pp. 141–144, 2016, doi: 10.1109/DeSE.2015.25.

[19] M. M. Zahid, K. Abid, A. Rehman, M. Fuzail, and N. Aslam, “An efficient machine learning approach for plagiarism detection in text documents,” *Journal of Computing & Biomedical Informatics*, vol. 4, no. 2, pp. 241–248, 2023.

[20] P. Gupta, A. B. -Cedeño, and P. Rosso, “Cross-language high similarity search using a conceptual thesaurus,” *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics,* vol. 7488, Springer: Berlin Heidelberg, pp. 67–75, 2012, doi: 10.1007/978-3-642-33247-0_8.

[21] J. P. Wahle, T. Ruas, T. Foltynek, N. Meuschke, and B. Gipp, “Identifying machine-paraphrased plagiarism.” *Information for a Better World: Shaping the Global Future*, 2022, vol. 13192, Springer, Cham, doi: 10.1007/978-3-030-96957-8_34.

[22] M. Potthast, B. Stein, A. Eiselt, A. B. -Cedeño, and P. Rosso, “PAN plagiarism corpus 2011 (PAN-PC-11),” *Zenodo,* 2011. [Online]. Available: https://zenodo.org/records/3250095

[23] R. Steinberger *et al.*, “The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages,” *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pp. 2142–2147, 2006.

[24] P. Koehn, “Europarl : a parallel corpus for statistical machine translation,” in *Proceedings of Machine Translation Summit X: Papers*, Phuket, Thailand, 2005, pp. 79-86.

[25] A. B. -Cedeño, C. E. -Bonet, J. Boldoba, and L. Màrquez, “A factory of comparable corpora from Wikipedia,” *8th Workshop on Building and Using Comparable Corpora, BUCC 2015 - co-located with 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2015,* Association for Computational Linguistics, pp. 3–13, 2015, doi: 10.18653/v1/w15-3402.

[26] J. Ferrero, L. Besacier, D. Schwab, and F. Agnès, “Deep investigation of cross-language plagiarism detection methods,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 6–15, 2017, doi: 10.18653/v1/w17-2502.

[27] X. Liu, K. Duh, L. Liu, and J. Gao, “Very deep transformers for neural machine translation,” *arXiv-Computer Science*, pp. 1-7, 2020, doi: 10.48550/arXiv.2008.07772.

[28] Y. Matrane, F. Benabbou, and N. Sael, “Sentiment analysis through word embedding using AraBERT: Moroccan dialect use case,” *2021 International Conference on Digital Age and Technological Advances for Sustainable Development, ICDATA 2021*. IEEE, pp. 80–87, 2021, doi: 10.1109/ICDATA52997.2021.00024.

[29] Z. Ellaky, F. Benabbou, and S. Ouahabi, “Systematic literature review of social media bots detection systems,” *Journal of King Saud University - Computer and Information Sciences,* vol. 35, no. 5, 2023, doi: 10.1016/j.jksuci.2023.04.004.

[30] N. Patwardhan, S. Marrone, and C. Sansone, “Transformers in the real world: a survey on NLP APplications,” *Information*, vol. 14, no. 4, 2023, doi: 10.3390/info14040242.

[31] Y. Liu *et al.*, “Multilingual denoising pre-training for neural machine translation,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020, doi: 10.1162/tacl_a_00343.

[32] J. Bromley *et al.*, “Signature verification using a ‘siamese’ time delay neural network,” in *Advances in Pattern Recognition Systems Using Neural Network Technologies*, pp. 25–44, 1994, doi: 10.1142/9789812797926_0003.

BIOGRAPHIES OF AUTHORS

Chaimaa Bouaine graduated with a Master's degree in Big Data and Data Science option Big Data from Hassan II University in Casablanca, Morocco in 2021. Currently, she is preparing her Ph.D. in Computer Science at the Laboratory of Information Processing and Modeling (LTIM) at the Faculty of Science Ben M’SIK. Her research interest is cross-language plagiarism detection including natural language preprocessing, machine learning, and deep learning. She can be contacted at email: chaimaa.bouaine-etu@etu.univh2c.ma.

Faouzia Benabbou is professor of Computer Science in the Department of Mathematics and Computer Science at the Ben M'Sick Faculty of Science, Hassan II University of Casablanca. She received his Ph.D. in Computer Science from the Faculty of Sciences, University Mohamed V, Morocco, in 1997. She is a member of the Computer Science and Modeling Laboratory and head of the team of Cloud Computing, Network and Systems Engineering (ICCNSE) team. Her research interests include cloud computing, data mining, machine learning and natural language processing. She can be contacted at email: faouzia.benabbou@univh2c.ma.