Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

AUTOMATIC PLAGIARISM DETECTION AND EXTRACTION IN A MULTILINGUAL: A CRITICAL STUDY AND COMPARISON

NEHA N. CHAUBEY,
Student, Electronics and Communication Department, Dharmsinh Desai University, Gujarat, India.

NIRBHAY KUMAR CHAUBEY,
Dean, Department of Computer Science, Ganpat University, Gujarat, India.
Corresponding Author

Abstract - The effectiveness of plagiarism detection is challenging because of the large quantity of accessible words of multiple languages on the internet. Plagiarism arises in various levels of complication extending from the original resource data to the concise text. Detection of plagiarism contents from one language to multilingual is one of the prime concerns. In the previous studies, extensive research works are presented to detect plagiarism contents to monolinguals. Although, less reliable research work takes place to detect multilingual plagiarism wherein one writer's work contents plagiarized by another writer. This is a major challenge for the researchers, academic institutes and research organizations and conference organizers to check the authenticity of the work and it has been gaining more focus in the research area in recent years. This paper extensively reviews the state-of-the-art various plagiarism detection and extraction techniques for the monolingual, bilingual and multiple languages and comprises the discussion. Moreover, benefits and limitations of the various deep learning based multiple language plagiarism detection techniques with the supported languages are reviewed. Consequently, this paper highlights some better techniques for plagiarism detection depending on machine learning techniques and deep learning based solutions.

Index Terms –Plagiarism, Cross-language plagiarism detection, Cross-language dataset, Natural Language Processing (NLP) Techniques.

INTRODUCTION

Plagiarism is the act of stealing someone else's work and trying to misrepresent as your own. This applies on all terms like papers, thoughts, ideas, songs etc. In other words, the dishonorable practice accompanied by a person is entitled as plagiarism. Nowadays, plagiarism is an increasingly extensive and growing issue in the academic world [1]. Moreover, detection of plagiarism is difficult because of the different types. Plagiarism is generally categorized into seven types. They are, rephrasing a content without correct citation, Mosaic plagiarism where data is collected from various sources and combined into one, copy & paste without the proper citation, incorrect citation, appropriating others whole work, citing non-existing work and self-plagiarism where one submits their published work as although it were new one [2]. Plagiarism detection in text documents is a succeeding subarea in research text processing. Nowadays plagiarism tends to be an aggregate challenge in society. Academia and population

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

industries are distressed due to plagiarism. Plagiarism detection services have been utilized to retain academic integrity, protect intellectual property and educational institutions [3].

Plagiarism is emphasized as a moral evil doing or legal offence and this can be detected with the inception of detection approaches. Detection technique is a progression of ruling plagiarized documents or texts. Plagiarism can be categorized into external plagiarism and intrinsic plagiarism. The comparison of mistrustful documents in contradiction of original documents is designated as external plagiarism. On the other hand, the plagiarism text within the paper without the ingress of the original test is termed as intrinsic plagiarism. Extrinsic plagiarism depends on reference collection and in other cases intrinsic plagiarism does not depend on reference collection to detect plagiarism [4]. In semantic plagiarism the contents are get rehabilitated but the innovative remains same so it is termed as idea plagiarism (dangerous plagiarism). Both the lexical and semantic based features are most probably used in plagiarism detection gauge the resemblance sandwiched between two texts [5].

Before detection accuracy, the input text such as tokenization, stop words removal, post tagging and text segmentation are characterized by numerous NLP techniques [6]. The accuracy of plagiarism detection gets enhanced by consolidating natural language processing (NLP) techniques. NLP technique ensures the availability of intended text [7]. Document source comparison, Manual search of characteristic phrases and stylometry are the three main categories of detection techniques [8].

Similarity in a transcript article originates in the regular text, tables, flow diagram, figure captions, and code. This similarity is accomplished by merely replicating, summarizing or through obscuring the word. The attributes in plagiarism normalized into three phases namely lexical translations, syntactic conversions and semantic changes [9]. In lexical conversation prevailing words or contents are swapped by synonyms or either adding or deleting. Syntactic conversions are executed by reordering the sentence construction, conversion of active to passive or passive to active elements and reconstruction of syntax format. Combination of both is emphasized as semantic conversions. A wide range of NLP techniques are introduced by researchers to examine, describe and interpret various languages spontaneously [10]. In NLP, language resources play two roles initially, large-scale interpreted corpora to motivate statistical NLP techniques and secondly evaluate in contradiction of a gold-standard for test assemblage.

The NLP resources are recognized based on assessment Map [11]. The NLP methods, Naïve Bayes, N gram are appropriate to find plagiarism in languages [12]. The plagiarism detectors can be branded as two systems such as monolingual and multilingual (cross language). Both the suspicious and source documents are similar in monolingual systems. But in cross language both are dissimilar [13].

Recently, multilingual text processing captures more attention in a multilingual environment. Multilingual plagiarism detection primarily smears to ruling the contents or texts hackneyed from dissimilar languages. It permits to recognize plagiarized

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

documents written from other sources [14]. Multilingual Plagiarism checking broaches the unrecognized reclaim of a text and it concerns the transformation from one natural language to another deprived of appropriate position to the inventive source but the concepts endure unchanged. Therefore revolution in syntax and semantic text necessitates a deep and concentrated processing. The detection of plagiarized content into distressful documents and their resultant fragments into source documents is a task of multilingual plagiarism. The detection method comprises five stages namely language normalization, recovery of candidate documents, classifier training, plagiarism analysis, and post-processing [15]. Here we have mainly focused on two languages namely Persian-English and Arabic-English in a multilingual plagiarism detection.

2.0 REVIEW ON AUTOMATIC DETECTION AND EXTRACTION OF PLAGIARISM

In this section, numerous techniques for the monolingual plagiarism detection, bilingual plagiarism detection and multilingual plagiarism detection are discussed. Checking multilingual plagiarism is the necessity of time. This section investigates some developed techniques found for plagiarism effectively for the various languages. The monolingual, bilingual and multilingual similarity identification techniques are described in subsections.

2.1 MONOLINGUAL PLAGIARISM DETECTION

Monolingual plagiarism detection refers to the prediction of plagiarized content between the monolingual documents. The monolingual documents cover only the contents copied from a single language rather than bothering too many languages. Due to the vast and exponential growth of the internet and the publication of research works, replication of the internet sources also increased a lot. Most importantly, the copying of contents from different languages have increased in recent times, which posts challenges in identification. Monolingual plagiarism detection paves way for efficient recognition of copied contents from any particular language into the same language. This kind of stealing of information without adding any citations must be prohibited. Keeping this in mind, researchers made attempts to efficiently recognize the plagiarized content in the monolingual environment.

The unauthorized usage of content without proper citations is considered as an academic threat that requires proper authenticity. To authenticate the plagiarized content, efficient strategies are required to be built with high accuracy. Normal copy-paste acts can be easily detected using certain detection tools. But the high-tech way of copying such as linguistic translations, paraphrasing and summarizations are difficult to recognize as the entire texts will be altered without using the similar words from the original document. The completely altered documents are the ones that are considered as suspicious documents by the plagiarism detection systems. These suspicious documents hold contents that are completely reworded with appropriate synonyms and the detectors attempt to identify the documents through the evaluation of synonyms of

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

the documents. The detectors are classified as intrinsic and extrinsic based on the type of comparisons they make to identify the plagiarized content. Intrinsic types focus on the deviations in the writing styles in a single document and extrinsic types compare the suspicious documents with a huge number of related documents sourced from the internet [16].

The identification of plagiarized content in the monolingual environment was highly required to avoid such misconducts. Multiple research works focused on detecting plagiarism from monolingual documents with different research viewpoints. Among all such researches, Asghari et al. [17] focused on detecting plagiarism in the Persian language based on the Persian corpus for plagiarism detection. The technique used HAMTA corpus for Persian plagiarism detection with manual paraphrase of contents for efficient detection. The technique also suggested crowd workers to manually paraphrase the documents into fragments. The degree of correlation was identified between the proposed and the existing plagiarism detection corpuses from experimentation.

There are certain languages that are highly complicated and require highly efficient and accurate techniques to resolve the problems related to plagiarism. One among those languages was Arabic where the research area stays weak due to its complicated tendency. Few researchers still focused on detecting plagiarism in the Arabic language. Mahmoud and Zrigui et al. [18] presented a strategy capable of generating paraphrased corpus automatically, where the corpus was predefined with almost all the vocabularies related to particular words. The corpus stayed as a core document to identify the plagiarized content in the documents in Arabic language. The index words identified were replaced with the mostly used words and the vocabularies are traced for all the words covering the synonyms and their semantic relationships. The experiment was carried out with a focus of identifying the closest words in semantic order through the variation in the dimensions of the vectors and sizes of the windows.

Detecting plagiarism combines two major steps such as retrieval of suspicious candidates and detailed analysis. Retrieving the suspicious candidate was a step where the computation of detection will be performed for all the suspicious documents. The computation process for detailed analysis entails complexity and is highly time consuming as it has to evaluate a huge count of documents for detecting plagiarism. An article by Ehsan and Shakery et al. [19] provided a system to recognize plagiarism using the Persian corpus of plagiarism. The system worked on detailed analysis also named pairwise document analysis centered on a vector space model. The system was modelled into three main blocks namely seeding, match integration and extrication filtering. The main advantage was that the approach had the capability of detecting plagiarism for any different linguistics. The evaluation was carried out for the Persian corpus to prove its excellence.

Improving the scalability and accuracy of the plagiarism detection techniques was still a challenge due to the unmanageable information available online. Since the information

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

was copied from other languages, managing and detecting plagiarism for diverse linguistic documents was a difficult task. Therefore, the detecting systems should be robust enough to trace any kind of copying like summarization, paraphrasing, altering, etc., and should be alert in identifying the varied types of obfuscation in the plagiarized documents.

An approach built on the basis of word embedding was the approach found by Gharavi et al. [20] where the text embedding vectors were used to identify the similarity between diverse documents. The syntactic and semantic information of the document were aligned using an aggregation function with the word vector for document representation. The document representation was compared with the suspicious document and the sentences posting high similarity were traced out. After tracing, the sentences were filtered and merged through online and offline tuning. The obfuscations in the plagiarized documents were efficiently traced out through that technique.
Plagiarized content was the work stolen from other's work posting challenges in identification. The structure and the semantics will be changed while translation and the meaning remain the same. It leads to problems because the idea of the original author was stolen without any credits and was used as a work done by some other author. Even if two different authors choose the same title for expanding the content, the writing style must be dissimilar to some degree other than the cited portions. Direct or indirect citing by any means or by quoting or referencing the author's details, the citation can be known and the plagiarism in that part was permissible to a particular extent. But the uncited part should not hold any plagiarism as it ruins the ethical behaviour of the documents available online [21].
Mohtaj et al. [22] suggested an approach to create a monolingual corpus for plagiarism detection with an aid of task alignment for PAN 2015 competition. The major focus was the identification of obfuscations between the documents and the technique employed two unique methods called artificial and simulated obfuscations for the creation of plagiarism cases.
The artificial method includes synonym substitution, changing the order randomly, preserving the changed order and addition or deletion whereas the simulated method constructed the plagiarism cases with scores obtained through sentence pairing. Though the corpus was constructed in English language, it was suggested to be used for other languages also for plagiarism detection.

2.2 BILINGUAL PLAGIARISM DETECTION

Bilingual plagiarism detection refers to the identification of plagiarism in a document through evaluation of two different linguistic documents. The suspicious documents might include contents copied from different languages and the copied contents are very difficult to recognize. Some documents might be copied through translation of a single document where the unilingual or monolingual detector works well for identifying

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

such cases. But when it comes to two different languages, an efficient bilingual plagiarism detector has to be introduced to trace the copy. The bilingual detector evaluates two different linguistic corpuses for every suspicious document to identify the similarity. Most importantly, any bilingual plagiarism detector should satisfy certain constraints for acceptance. The detector should possess the capability to handle two different linguistic thesauruses each with equal efficiencies. Both the testing and training parts should follow the same approaches or algorithms without any changes for evaluation. This constraint was ideal as it determines the real accuracy of the developed system. After identification of relevance, the document has to be sorted in descending order where the document in the beginning shows the highest similarity.

Researches related to bilingual plagiarism detection were in progress and many researchers were focusing on finding the similarities between the documents of complicated languages. Most of the copied documents were from other natural languages to English language. An effort was made by Arefin et al. [23] to identify the similarities between the electronic Bangla and English documents. The system followed two dissimilar techniques such as analysing the individual contents present in the document such as stop word removal, keyword extraction, checking the synonym and bilingual translation and the second method involved performing statistical analysis for the documents.

Upon evaluation, the technique proved to accurately identify the difference between the Bangla and English documents. Moreover, the technique was also capable of finding the similarities from other linguistic documents with slight modifications in the thesaurus and storage format. The only flaw that required improvement in the above technique was that it used the in-house database of the documents which posted the requirement of scanning a document before detecting the plagiarism.

Due to the rapid growth of the internet and the contents, there was a rapid increase in plagiarism between the documents that were made intentionally, sometimes even unintentionally, both of which were considered unethical. Bilingual methods mostly follow statistical methods for identifying the plagiarism and two diverse unrelated languages will be considered for evaluation. Arefin et al. [24] introduced a bilingual plagiarism detector to identify the plagiarism between the Bangla and English documents. A statistical method followed by individual content detection were entailed as the major methods in the suggested technique. The technique also had the capability to identify the similarities between the translations of Bangla to English and vice versa along with the capability of identifying the similarities between the same linguistic documents. The system was demonstrated to be effective for prediction of plagiarism in the Bangla and English documents but certain modifications were needed to use the system for other languages.

The plagiarism detection corpus for the Persian and English languages was constructed by Asghari et al. [25] with an aim of constructing a text alignment corpus for the PAN

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

2015 competition. This corpus was the first known bilingual corpuses to be constructed. The corpus was created using the contents from Wikipedia articles along with the aligned parallel corpus constructed from the Persian and English languages. The similarity between the documents was identified through providing a similarity score for the documents between 0 and 1. An obfuscation strategy was introduced based on the similarity degree with a capability of adjusting it between the plagiarized documents. Clustering and fragments obfuscation was done for the documents for the construction of the corpus with high similarity identifications.

The area of research in plagiarism has grown a lot but there was an absence of attention in the field of bilingual plagiarism prediction. Certain methods work on evaluating the detection techniques with other linguistic corpuses to identify the excellence in performance. One such method was the Chinese-English bilingual plagiarism detection method introduced by Chen and Vines et al. [26]. It was a multi-querying method that enclosed certain queries to find out the similarity in bilingual documents. Several other methods related to candidate document retrieval were introduced and four query methods were compared with the presented method.

Based on the queries submitted, the document will be retrieved and the queries possess certain keywords for identification. The results of the above mentioned technique suggested that the query to be submitted should possess 50% keywords for exact identification of the document.

The bilingual corpus can be used not only to evaluate bilingual plagiarisms but also to identify plagiarisms in a multilingual environment. An author named Asghari et al. [27] introduced a Persian-English corpus for the identification of plagiarisms between cross-language documents. The constructed bilingual corpus possessed seven different types of obfuscations and the corpus was named as HAMTA-CL. The method used the word embedding model to detect the plagiarisms between the documents of diverse languages. The corpus was constructed through insertion of certain fragments of text into the suspicious document for plagiarism identification. The articles obtained through Wikipedia stayed as the prime source behind the building of the HAMTA-CL corpus. The pre-processing was utilized for normalization of Wikipedia documents and a total count of 1904 documents with varied lengths were used for the corpus construction. The experiments carried out proved the efficiency of the constructed corpus as well the efficiency of the suggested plagiarism detection technique.

Shiraz and Yaghmaee et al. [28] presented a solution for bilingual plagiarisms through the capability of detecting plagiarism automatically between the English and Persian languages. The method was highly accurate and automatic detection was achieved through the use of a vector space model (VSM). The method proved to show high accuracy and reliability in plagiarism detection under experimentation.

The technique was evaluated using the Persian texts expanded into two sets as

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

plagiarized and non-plagiarized with morphological exploration and without morphological exploration. The results suggested that the technique was especially suitable for the identification of plagiarized contents between the Persian and English languages.

The semantic analysis plays a role in the identification of plagiarized content from any kind of documents irrespective of the translations from different linguistics. Ratna et al. [29] introduced a technique to identify the plagiarism between Indonesian and English languages. The technique applied semantic analysis (LSA) and vector quantization (LVQ) to identify the similarities of the two languages possessing different syntax. Indonesian language was not considered mostly by any other researchers due to the lack of NLP for that language. Hence LSA was chosen for plagiarism detection as it only focuses on the presence of words regardless of the grammar. Finally, the LVQ classifier was used to classify the data depending on the distance between the input and output. The system performed well under experiments and showed higher recall value than any other systems compared with it.

2.3 MULTI-LINGUAL PLAGIARISM DETECTION AND EXTRACTION

Multilingual plagiarism detection plays a significant role in the identification of copied texts that belong to other authors which are in different languages. The documents published by other authors in different languages might be converted into a different language and used as content by another author. This is a serious ethical delinquency and also involves problems like the spread of wrong information without the knowledge about the original source. The reuse of content with translation is unbearable as most of the detection techniques are not capable of finding the copied content behind language barriers. There are surplus documents available on the internet that have opportunities of sourcing error contents and hence the spoilage of original source may encounter which is immoral. To avoid such critical issues, multilingual plagiarism detection plays a key role through sorting out of the contents that are translated from other natural languages.

Due to the drastic growth of the internet, the contents present in it has also increased a lot which lead to a dramatic increase in plagiarism. Multilingual plagiarism detection mainly applies to finding out of the contents copied from diverse languages. But it also has the capability to find out the contents sourced from the internet between similar languages. In case of multilingual plagiarism, the text format is alone changed whereas the contents belonging to some other author remains unchanged. Normal plagiarism detectors are not capable of reading such changes as it requires a deep processing with complicated systems. Moreover, the system should check through all the available documents on the internet and should possess the capability of managing huge quantities of data. Managing vast contents from the internet include varied complexities such as time complexity, efficiency, etc. All these issues must be taken into consideration while finding a solution to detect multilingual plagiarism.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

Meysam Roostaee et al. [30] presented an innovative scheme that depends on multilingual word embeddings. The goal of utilizing syntactic and semantic data to arrange the similarity contents from the basis and concerned file precisely. Initially, a vector space framework works on multiple language word embeddings dependent on a methodology of local weighting. It was utilized to extricate a least group of extremely possible pairs instead of analysing all possible pairs of data. The presented approach step comprises a dynamic development methodology to shield additional candidate pairs targeting at enhancing the framework's recall. This was surveyed through a more precise technique that observes the language pairs in the sentence level utilizing a graph of words demonstration of words.
Mohran et al. [31] presented a multi-language plagiarism detection methodology to examine and advance the detection. Simplest and effective presented multilingual supportive similarity predictor with the ease of one press to identify the text plagiarism. The presented procedure was developed with an intelligent framework that was able to acquire, modify and adjust based on the input text and create a fast search for the words on the limited and the accessible repository and connect the word of the document with the similar corresponding word anywhere identified.

J´er´emy Ferrero et al. [32] presented a deep investigation on approaches for identifying multilingual plagiarism on an openly accessible database. The considered database consists of French, English and Spanish language texts. Assessment procedure is utilized and the outcomes expose that if the database is adequately abundant in one language. Moreover, the technique outputs are effective and returns the equivalent for another number of languages also. It proves that the technique is better in a specific language, it is better in other languages also.

Jeremy Ferrero et al. [33] presented a methodology for the assessment of cross language document plagiarism identification. They improved a previous corpora and the drawbacks of that work and they utilized different collected assets to overwhelm the limitations. The developed database is based on multiple languages comprising cross linguistic arrangement for various granularities and depends on both parallel and similar corpora and comprises human and machine interpreted texts. Furthermore, it comprises content written with the different writers. With the attained database, they conduct an organized and demanding assessment of numerous multiple language based plagiarism identification techniques.

Marat Rakhmatullayev et al. [34] presented a technique to find plagiarism amongst more languages. Here, a dictionary dependent machine conversion technique was utilized for the conversion of terms. Cross language based identification of plagiarism process comprises two steps are paper indexing and recognition procedure. The two stages were described and the framework of the plagiarism detection with a multi

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

number of languages is presented with the developed methodology.

Meysam Roostaee et al. [35] developed an innovative technique to recover the similarity contents through multiple languages. A synthesis of theoretical and keyword dependent techniques are utilized for the similarity prediction. In the presented methodology, a dynamic factor was considered to associate the outcomes of conceptual and bag of words prototypes. The introduced technique cross language similarity was done with the combination of German-English-Spanish languages. But the previous compared techniques were used with only two combined languages.

Marc Franco-Salvador et al. [36] presented a Cross language similarity detection on knowledge graph based illustrations of multiple languages. The presented hybrid techniques evaluate the semantic resemblance among words in dissimilar languages. The presented methodology hybridizes the demonstrations at developing their similarity in attaining various features of cross language plagiarism.

Moreover, they analyzed a continuous content arrangement dependent plagiarism examination with a novel procedure to evaluate the resemblance among text fragments. They analyzed the developed technique with the previous numerous schemes in the analysis of similarity. The multiple language database comprises of Spanish-English (ES-EN) and German-English (DE-EN) for plagiarism recognition.

Marc Franco-Salvador et al. [37] presented an innovative graph centered methodology that utilizes a multilingual semantic model to analyze the content's similarity in numerous varying languages.

To examine the presented methodology, they utilized the German, English and Spanish languages for the similarity identification with the available PAN-PC'11 corpus. They compared the attained outcomes with the previously available techniques. Exploratory outcomes designate that the presented graph based methodology was a correct alternative for cross-language similarity prediction.

Marc Franco-Salvador et al. [38] developed a multiple language semantic framework with organized investigation of Knowledge Graph Examination. The presented scheme signifies content similarity through a knowledge graph. It was a language autonomous content prototype. They investigated the assistance to cross language similarity identification of the various features enclosed through knowledge based graphs. Moreover, they investigated related ideas and that association on detecting plagiarism. Lastly, an innovative weighting strategy was introduced for contrasting the knowledge graph through dispersed demonstrations of concepts. Here, BabelNet multilingual semantic network was utilized.

Safi-Esfahani et al. [39] developed a methodology based on semantic procedure for the multiple language similarity prediction. The presented work predicts the similarity of at times mentioned as the multiple language similarity. In that, content novelists combine a transformation with their contents. Centered on the mono language similarity prediction technique, the developed work eventually proposes to discover a methodology to identify the cross language plagiarism. A structure named multiple language plagiarism

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

detection (MLPD) was accessible for the similarity examination with the better objective based prediction of similarity. Here, English was considered as a referenced language and Persian and other languages were back interpreted utilizing the conversion tools. The information utilized for MLPD evaluation was attained through the English Persian Mizan parallel corpus. The benefits and limitations of the various multiple language plagiarism detection techniques are described in table 1.

Table 1: Brevity of Various plagiarism detection and extraction techniques in multilingual

Authors	Methods	Benefits	Limitations
MeysamRoostaee et al. [29]	Multilingual similarity detection with dictionary and a local weighting procedure	Effectually identifies the similarity through selecting the finest translation of every word instead of using all the translations Considerably reliable.	Needs to improve the illustration via exploiting a Weighting technique.
Mohran et al. [30]	Intelligent Multi-Language Plagiarism Detection framework	Improved features for contents searching and provides optimized results	Machine learning approaches are needed to improve the prediction.
J´er´emy Ferrero et al. [31]	Deep examination on similarity identification	The optimized technique on a on a specific corpus is proficiently utilized on other corpus.	Learning based techniques are not provided.
Marat Rakhmatullayev, et al. [33]	Dictionary based machine translation method	Provides better similarity detection	Performance can improve more.
MeysamRoostaee et al. [34]	Fusion method with conceptual and key word based prototypes	Enhances the similarity prediction accuracy Content based similarity attains highest score. High proficiency procedure	In future needs to utilize other type of knowledge bases for the analysis.
Marc Franco-Salvador et al. [35]	Knowledge graph based illustrations	Computational efficiency of the presented technique is improved.	Detection accuracy is not enhanced well.
Marc Franco-Salvador et al. [36]	Novel graph based scheme with multiple language semantic network	Presented methodology is the better alternative for the similarity cases of PAN-PC'11 corpus	Need of extraction of similarity with the identification
Safi-Esfahani et al. [38]	Semantic features based analysis	Attains highest accuracy.	Other translation methods are needed to improve the performance

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

Marc Franco-Salvador et al. [37]	Knowledge graph analysis	Achieves higher performance The presented technique is highly adequate	New techniques with knowledge graph is needed to improve the accuracy.

The table 1 depicts the various multilingual plagiarism detection techniques and their benefits and limitations of the investigated approaches were discussed and also reviewed the corpus utilized in the multilingual plagiarism detection techniques.

MULTILINGUAL PLAGIARISM DETECTION WITH DEEP LEARNING

El Mostafa Hambi and Faouzia Benabbou [42] developed a plagiarism identification structure depends on the combinations of different deep learning representations were, Convolutional Neural Network (CNN), Siamese Long Short-term Memory (SLSTM) and Doc2vec. The presented approach utilizes three layers of pre-processing layers consisting of insertions of word, Learning and prediction layer. To validate the presented framework, different plagiarism identification tools from the research field are considered depending on a set of features.

Hananeezzikouri et al. [43] presented an improved fuzzy semantic dependent plagiarism scheme for examining and matching texts in according to the WordNet lexical dataset, to identify plagiarism in texts interpreted from or to Arabic, a pre-processing stage was necessary to proceed possible information for the fuzzy procedure. The presented methodology was analyzed with the texts of Arabic/English taking into deliberation the characteristics of the Arabic language. The presented scheme provides the automatic multilingual plagiarism identification through fuzzy semantic similarity.

Dima Suleiman et al. [44] presented deep learning with word2vec technique to predict the semantic resemblance among words in Arabic language to find the similarity. Word2vec was a deep learning methodology that was utilized to signify texts as a vector of features with highest accuracy. The quality of vector illustration depends on the superiority of chosen corpus utilized in the training stage. The presented framework utilized the OSAC corpus for preparation of the word2vec network. Furthermore, cosine similarity was utilized to calculate the plagiarism amongst the words vector.

El Mostafa Hambi, Faouzia Benabbou [45] presented an analysis that depends on a measures of vector demonstration technique, Level Treatment and plagiarism techniques. The outcomes of the presented study was that the utmost of investigates depended on world granularity and utilized the word2vec methodology for demonstrating the word vector. In some cases, it was not appropriate to preserve the meaning of all paragraphs. Every analyzed methodology was providing better outcomes

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

and there was a chance to improve the techniques more. Plagiarism detection can be used online for making the job easy instead of lacking more required storage space for large tools. Online plagiarism detecting tools also works efficiently in many cases. Also, these tools are easy to handle and may not post high cost for any number of uses.
Alabbas et al. [46] attempted to introduce an efficient online plagiarism detection system which was highly easy and simple to use. The major focus was to identify the plagiarism among the Arabic documents and it involved only pasting of the content into the tool used for detection. The tool made use of Google, Yandex SERP and Bing to traverse the presence of pasted content. The search APIs of the abovementioned search engines paved the way to collect the plagiarised content. The experiments proved that the technique introduced was efficient to identify plagiarism among different documents online.

Salha Alzahrani and Hanan Aljuaid [47] developed a methodology to extricate the combined features and their hybridization from the multiple language script and analyze the relations about the similarity. The proposed work concentrates to examine the subsequent stages termed as cross language semantic text similarity (CL-STS) and plagiarism detection (PD) utilizing deep neural networks (DNN). The presented combination of technique for the identification of the similarity among multiple languages was not utilized in the previous works and it was newer and effective than the other previous schemes. Moreover, the presented diverse neural network frameworks provide the solution for the PD for example binary classification or deeper categorization. Moreover, DNN was utilized as an objective function to evaluate the semantic implications for CL-STS.
El Mostafa HAMBI and Faouzia Benabbou [48] introduced a framework of plagiarism identification depending on Deep Learning Algorithms. The presented methodology considers the difficulties faced in similarity identification of concepts like: meaning losses otherwise the challenges in prediction of semantic resemblance among multiple languages. Consequently, the presented framework comprises of utilizing in a doc2vec to attain a vector illustration of every word of a text and utilizes the siamese LSTM to learn the introduced framework that pair of documents is comparable and lastly CNN model was utilized to categorize the various kinds of plagiarism. The benefits and limitations of the various multiple language plagiarism detection techniques are described in table 2.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

Table 2: Deep Learning based multiple plagiarism detection and extraction techniques in multilingual

s	Methods	Benefits	Limitations
El Mostafa Hambi and Faouzia Benabbou [42]	Combined deep learning models of Doc2vec, SLSTM and CNN	The developed scheme detects the similarity as well as the classes of the similarity. Computational time is decreased.	Needs to consider the accuracy improvements
Hananeezzikouri et al. [43]	Fuzzy semantic dependent similarity framework	Provides improved accuracy.	Extended methods are needed to improve the performance.
Dima Suleiman et al. [44]	Deep learning framework	Provides high precision output and accurate identification of similarity.	Computational complexity of the framework is high.
El Mostafa Hambi and Faouzia Benabbou [45]	Deep learning neural network	The deep learning framework is reliable	There is an possibility of improving the performance
Salha Alzahrani and Hanan Aljuaid [47]	Rich Semantic Features and Deep Neural Networks	Attains better accuracy and categorize the classes effectively. Combined methodologies are effective in the prediction of similarity	Computational complexity of the technique is high.
El Mostafa HAMBI and Faouzia Benabbou [48]	Deep learning techniques	Proficient in prediction of various types of similarity	More language datasets are needed for the analysis to progress the presented framework as effective.

The table 2 signifies the review on various multilingual plagiarism predictions depending on deep learning. The benefits and limitations of the investigated approaches were discussed in this section accordingly. Moreover, the languages supported and the corpus utilized in the multilingual plagiarism detection techniques are reviewed.

EXISTING PLAGIARISM DETECTION TOOLS In this section, existing available plagiarism detection tools are discussed. Some the existing tools are described in

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

subsections,

The Turnitin tool: This detection framework permits users to check their contents and compare with web data and other source documents available on the internet. For every submission, similarity is found in percentage. This tool uses the matching algorithm for measuring the similarity among the data.

The ithenticate tool: It is an online academic based plagiarism detection tool for the research scholars. This tool is used by the publishers, researchers, and authors. This tool uses their own dataset for predicting plagiarism. Their dataset comprises a number of documents including books, articles and newspapers. For plagiarism prediction, this tool supports more than 30 languages like Arabian, Russian, English etc.

The Urkund tool: This plagiarism detection tool is used for preventing the data stealing in web data sources. This finds the similarity of the submitted document with the several data sources. It results in the similarity rate of the submitted document.

The plagiarism detection.org tool: This plagiarism detection tool is an online tool mostly utilized by the school students and teachers. The major advantage of this tool is supporting all the languages. Moreover, it uses n-gram methodology for plagiarism detection.

The PlagAware tool: This tool uses a traditional approach for plagiarism detection. It can support only a few languages: German, English and Japanese.

These are some plagiarism detection tools available in the existing. But compared to the tools mentioned in this review, plagiarism detection systems using deep learning [43] based techniques will enhance the different functions of plagiarism detection. Moreover. It can enhance the accuracy of plagiarism prediction and also can minimize the response time of prediction.

RESULTS DISCUSSION

In this section performance analysis of various existing approaches in regards to evaluation metrics are mentioned. Here, results of various plagiarism detection techniques in respective of different evaluation metrics are given in table 3.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

Table 3: Performance analysis of various existing approaches and evaluation metrics

Techniques used in plagiarism detection	Recall (%)	Precision (%)	F1-score (%)
Cross-lingual bag of concepts [36]	60.41	40.06	48.17
T+MA-BOW [36]	77.75	51.36	61.85
Fusion approach [36]	81.05	60.88	69.53
Fine grained plagiarism detection [31]	94.96	26.2	-
Two level matching technique [31]	97.46	55.6	-
Latent semantic analysis [40]	76	67	61
Smith waterman algorithm [40]	76	23	77

Table 3 provides the different techniques performance evaluations in terms of precision, recall, F1-score. From the analysed approaches, fine grained plagiarism detection provides better improvement than the other mentioned techniques. Moreover, the accuracy performance with different techniques are mentioned in table 4.

Table 4: Comparison on accuracy performance

Techniques used in plagiarism detection	Accuracy (%)
Support vector machine (SVM) [46]	92.83
Deep learning [48]	95
Logistic regression [46]	92.88

Table 4 provides the performance of few important approaches used in plagiarism detection. The deep learning based approach attains improved performance than the other techniques. Hence, deep learning based techniques with hybrid approaches are suggested in future works for effective plagiarism detection with better performance.

3. CORPUS AVAILABLE FOR MULTILINGUAL PLAGIARISM DETECTION

This section presents a summary of corpora available for multiple language plagiarism identification. The corpus available for multiple language plagiarism detection is mentioned in table 3.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

Table 5: Corpus available for cross language plagiarism detection.

Authors	Languages supported	Corpus available
El Mostafa Hambi and Faouzia Benabbou[40]	Multiple languages	PAN-PC-11
Dima Suleiman et al. [42]	Six languages comprising Arabic and English etc.	OSAC corpus
El Mostafa Hambi and Faouzia Benabbou [43]	Natural languages	OSAC Arabic corpus, PAN 2016 etc.
Salha Alzahrani and Hanan Aljuaid [45]	71,910 Arabic English pairs	Standard benchmark corpus
El Mostafa HAMBI and Faouzia Benabbou [46]	Multilingual	PAN corpus
Meysam Roostaee et al. [29]	German, English, Spanish,	PAN-PC-11 and PAN-PC-12 corpus
Mohran et al. [30]	Foreign languages	100 different corpus consisting of short (200-300 words)
J´er´emy Ferrero et al. [31]	French, English and Spanish	Both parallel and comparable corpora (mix of Wikipedia, scientific conference
Marat Rakhmatullayev, et al. [33]	Multiple languages	-
MeysamRoostaee et al. [34]	German, English, Spanish	JRC-Acquis corpus
Marc Franco-Salvador et al. [35]	Spanish, English, German	PAN-PC-12 dataset
Marc Franco-Salvador et al. [36]	German, English, Spanish	PAN-PC'11 corpus
Safi-Esfahani et al. [38]	English Persian Mizan parallel languages	English-Persian Mizan parallel corpus
Marc Franco-Salvador et al. [37]	Spanish, English, German	PAN-PC-1030 and PAN-PC-11

4. DISCUSSION

This section describes the challenges faced in plagiarism detection are as follows: The efficiency of the plagiarism detection frameworks are very challenging because of multi-language textual data in the internet. The robust prediction of plagiarism with the developed techniques are highly complicated.

Maximizing the detection performance with precision and recall leads to lesser execution performance. Balancing all the performances are difficult with the developed techniques.

The developed techniques are not efficient for plagiarism checking in multiple

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

languages. Moreover, identification of similarity based on replacing words with their synonyms is difficult.
Challenges such as time complexity, database inclusion, and word sense disambiguation are not dealt with properly in the developed techniques.
NLP techniques perform well with simulated results while giving a lessened performance with artificial plagiarism. More language datasets are needed for the analysis to progress the presented framework as effective.
Computational complexity of the developed techniques are high and there is a possibility of improving the performance with the extended methods. Moreover, extended techniques are needed in the similarity extraction also.
The Deep learning based plagiarism detection techniques with the hybrid approaches can be further used for enhancing the plagiarism detection.

5. CONCLUSION

This paper presents a systemic state-of-the art critical review of different plagiarism detection techniques. Initially provided the brief discussion about the monolingual and multilingual plagiarism identification techniques. Afterwards, multilingual plagiarism detection with deep learning techniques are analyzed along with their benefits, limitations and the supported languages in the comparative table. Moreover, the availability of corpora in different language pairs are provided. This paper enables buddy researchers to make a choice for the automatic plagiarism detection and extraction on multilingual data with the improved deep learning and machine learning techniques. Among the analyzed techniques, deep learning based approaches are providing better improvements in plagiarism detection. In future work, multiple language plagiarism detection with hybrid machine learning techniques will be implemented to find the real time accuracy of plagiarized contents in cross language.

REFERENCES

Peters, Michael A., Liz Jackson, Ruyu Hung, Carl Mika, Rachel Anne Buchanan, Marek Tesar, Tina Besley et al. "The case for academic plagiarism education: A PESA Executive collective writing project." Educational Philosophy and Theory (2021): 1-24.

Al-Thwaib, Eman, Bassam H. Hammo, and Sane Yagi. "An academic Arabic corpus for plagiarism detection: design, construction and experimentation." International Journal of Educational Technology in Higher Education 17, no. 1 (2020): 1-26.

Samia, Zouaoui, and Rezeg Khaled. "Multi-Agents Indexing System (MAIS) for Plagiarism Detection." Journal of King Saud University-Computer and Information Sciences (2020).

Mahdavi, P.; Siadati, Z.; Yaghmaee, F.: Automatic external Persian plagiarism detection using vector space model. In: Proceedings of the 4th International Conference on Computer and Knowledge Engineering, ICCKE 2014, pp. 697–702 (2014)

Al-Shamery, EmanSalih, and HadeelQasemGheni. "Plagiarism detection using semantic analysis." Indian Journal of Science and Technology 9, no. 1 (2016).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

Ezzikouri, Hanane, Mohamed Oukessou, MadaniYouness, and Mohamed Erritali. "Fuzzy Semantic-Based Similarity and Big Data for Detecting Multilingual Plagiarism in Arabic Documents." In International Conference on Advanced Intelligent systems for Sustainable Development, pp. 159-169. Springer, Cham, 2018.

Aravind, K. S., BiradavoluShanmukh, Guru Sri Charan, and Chethan S. Nellikoppad. "A Survey of Cross-Lingual Plagiarism Detection using Natural Language Processing." International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653, Volume 8, 2020

Desai, Takshak, UditDeshmukh, Mihir Gandhi, and Lakshmi Kurup. "A hybrid approach for detection of plagiarism using natural language processing." In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, pp. 1-6. 2016.

Chong, M.Y.M.: A study on plagiarism detection and plagiarism direction identification using natural language processing techniques. Thesis Rep (2013)

El-Haj, Mahmoud, UdoKruschwitz, and Chris Fox. "Creating language resources for under-resourced languages: methodologies, and experiments with Arabic." Language Resources and Evaluation 49, no. 3 (2015): 549-580.

Calzolari, N., Soria, C., Gratta, R.D., Goggi, S., V.Q., Russo, I., Choukri, K., Mariani, J., &Piperidis, S. (2010). The LREC 2010 resource map. In The 7th international language resources and evaluation conference (LREC 2010), LREC 2010, Valletta, Malta, pp. 949–956.

Lamba, Harshall, and SharvariGovilkar. "A Survey on Plagiarism Detection Techniques for Indian Regional Languages." Int. J. Comput. Appl (2017).

Vani, K., & Gupta, D. (2017a). Text plagiarism classification using syntax based linguistic features. Expert Systems with Applications, 88, 448–464. https://doi. org/10.1016/j.eswa.2017.07.006.

Ceska, Zdenek, Michal Toman, and KarelJezek. "Multilingual plagiarism detection." In International Conference on Artificial Intelligence: Methodology, Systems, and Applications, pp. 83-92. Springer, Berlin, Heidelberg, 2008

Pereira, Rafael Corezola, Viviane P. Moreira, and RenataGalante. "A new approach for cross-language plagiarism analysis." In International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 15-26. Springer, Berlin, Heidelberg, 2010.

Chitra, A., and AnupriyaRajkumar. "Plagiarism detection using machine learning-based paraphrase recognizer." Journal of Intelligent Systems 25, no. 3 (2016): 351-359.

Asghari, Habibollah, OmidFatemi, SalarMohtaj, and HeshaamFaili. "A crowdsourcing approach to construct mono-lingual plagiarism detection corpus." International Journal on Digital Libraries (2020): 1-13.

Mahmoud, Adnen, and MounirZrigui. "Artificial method for building monolingual plagiarized Arabic corpus." Computación y Sistemas 22, no. 3 (2018): 767-776.

Ehsan, Nava, and AzadehShakery. "A Pairwise Document Analysis Approach for Monolingual Plagiarism Detection." In FIRE (Working Notes), pp. 145-148. 2016.

Gharavi, Erfaneh, HadiVeisi, and Paolo Rosso. "Scalable and language-independent embedding-based approach for plagiarism detection considering obfuscation type: no training phase." Neural Computing and Applications (2019): 1-15.

Alzahrani, Salha M., Naomie Salim, and Ajith Abraham. "Understanding plagiarism linguistic patterns,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

textual features, and detection methods." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, no. 2 (2011): 133-149.

Mohtaj, Salar, HabibollahAsghari, and VahidZarrabi. "Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus." In CLEF (Working Notes). 2015.

Arefin, Mohammad Shamsul, Yasuhiko Morimoto, and Mohammad Amir Sharif. "BAENPD: A Bilingual Plagiarism Detector." JCP 8, no. 5 (2013): 1145-1156.

Arefin, Mohammad Shamsul, Yasuhiko Morimoto, and Mohammad Amir Sharif. "Bilingual plagiarism detector." In 14th International Conference on Computer and Information Technology (ICCIT 2011), pp. 451-456. IEEE, 2011.

Asghari, Habibollah, KhadijehKhoshnava, OmidFatemi, and HeshaamFaili. "Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus." Notebook for PAN at CLEF (2015).

Chen, Hong Ye, and Phil Vines. "Multi Queries Methods of the Chinese-English Bilingual Plagiarism Detection." In Applied Mechanics and Materials, vol. 462, pp. 1158-1162. Trans Tech Publications Ltd, 2014.

Asghari, Habibollah, OmidFatemi, SalarMohtaj, HeshaamFaili, and Paolo Rosso. "On the use of word embedding for cross language plagiarism detection." Intelligent Data Analysis 23, no. 3 (2019): 661-680.

Shiraz, SorayaEnayati, and FarzinYaghmaee. "Introducing an Automated Technique for Bilingual Plagiarism detection of English-Persian Documents." (2014).

Ratna, AnakAgungPutri, Prima DewiPurnamasari, BomaAnantasatyaAdhi, F. AsthaEkadiyanto, Muhammad Salman, MardiyahMardiyah, and Darien Jonathan Winata. "Cross-language plagiarism detection system using latent semantic analysis and learning vector quantization." Algorithms 10, no. 2 (2017): 69.

Roostaee, Meysam, SeyedMostafaFakhrahmad, and Mohammad HadiSadreddini. "Cross-language text alignment: A proposed two-level matching scheme for plagiarism detection." Expert Systems with Applications 160 (2020): 113718.

Al-Bayed, Mohran H., and Samy S. Abu-Naser. "Intelligent Multi-Language Plagiarism Detection System." (2018).

Ferrero, Jérémy, Laurent Besacier, Didier Schwab, and Frédéric Agnes. "Deep investigation of cross-language plagiarism detection methods." arXiv preprint arXiv:1705.08828 (2017).

Ferrero, Jérémy, Frédéric Agnes, Laurent Besacier, and Didier Schwab. "A multilingual, multi-style and multi-granularity dataset for cross-language textual similarity detection." In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4162-4169, 2016.

Rakhmatullayev, M., J. Atadjanov, G. Lola, M. Yulduzxon, and A. Komila. "Cross-language plagiarism detection steps." International Journal of Scientific and Technology Research 9, no. 1 (2020): 3303-3308.

Roostaee, Meysam, Mohammad HadiSadreddini, and SeyedMostafaFakhrahmad. "An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes." Information Processing & Management 57, no. 2 (2020): 102150.

Franco-Salvador, Marc, Parth Gupta, Paolo Rosso, and Rafael E. Banchs. "Cross-language plagiarism detection over continuous-space-and knowledge graph-based representations of language." Knowledge-based systems 111 (2016): 87-99.

Franco-Salvador, Marc, Parth Gupta, and Paolo Rosso. "Cross-language plagiarism detection using a

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:01:2022
DOI 10.17605/OSF.IO/DWUK4

multilingual semantic network." pp. 710-713. Springer, Berlin, Heidelberg, 2013.

Franco-Salvador, Marc, Paolo Rosso, and Manuel Montes-y-Gómez. "A systematic study of knowledge graph analysis for cross-language plagiarism detection." Information Processing & Management 52, no. 4 (2016): 550-570.

Kurniawan, M. A., and K. Surendro. "Similarity measurement algorithms of writing and image for plagiarism on Facebook's social media." In IOP Conference Series: Materials Science and Engineering, vol. 403, no. 1, p. 012074. IOP Publishing, 2018.

Gharavi, Erfaneh, KayvanBijari, KiarashZahirnia, and HadiVeisi. "A Deep Learning Approach to Persian Plagiarism Detection." FIRE (Working Notes) 34 (2016): 154-159.

Safi-Esfahani, F., ShRakian, and M. H. Nadimi-Shahraki. "English-Persian Plagiarism Detection based on a Semantic Approach." Journal of AI and Data Mining 5, no. 2 (2017): 275-284.

El MostafaHambi, FaouziaBenabbou. "A New Online Plagiarism Detection System based on Deep Learning." (IJACSA) International Journal of Advanced Computer Science and Applications Vol. 11, No. 9, 2020.

Ezzikouri, Hanane, Mohamed Erritali, and Mohamed Oukessou. "Fuzzy-semantic similarity for automatic multilingual plagiarism detection." Int. J. Adv. Comput. Sci. Appl 8, no. 9 (2017): 86-90.

Suleiman, Dima, Arafat Awajan, and Nailah Al-Madi. "Deep learning based technique for Plagiarism detection in Arabic texts." ICTCS, pp. 216-222. IEEE, 2017.

El Mostafa, Hambi, and FaouziaBenabbou. "A deep learning based technique for plagiarism detection: a comparative study." IAES International Journal of Artificial Intelligence 9, no. 1 (2020): 81.

Alabbas, Maytham, Raidah S. Khudeyer, Mustafa Radif, and Hassan Khalid Hameed. "Online Multilingual Plagiarism Detection System Using Multi Search Engines." Journal of Southwest Jiaotong University 54, no. 6 (2019).

Alzahrani, Salha, and HananAljuaid. "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 (2020).

El Mostafa HAMBI and FaouziaBenabbou. "A Multi-Level Plagiarism Detection System Based on Deep Learning Algorithms" IJCSNS International Journal of Computer Science and Network Security, VOL.19 No.10, October 2019.