Arabic English Cross-Lingual Plagiarism Detection Based on Keyphrases Extraction, Monolingual and Machine Learning Approach

Mokhtar Al-suhaiqi1[1*] , Muneer A.S. Hazaa2[2] , Mohammed Albared3[3]

ABSTRACT

Due to rapid growth of research articles in various languages, cross-lingual plagiarism detection problem has received increasing interest in recent years. Cross-lingual plagiarism detection is more challenging task than monolingual plagiarism detection . This paper addresses the problem of cross-lingual plagiarism detection (CLPD) by proposing a method that combines keyphrases extraction, monolingual detection methods and machine learning approach.. The research methodology used in this study has facilitated to accomplish the objectives in terms of designing, developing, and implementing an efficient Arabic – English cross lingual plagiarism detection.

This paper empirically evaluates five different monolingual plagiarism detection methods namely i)N-Grams Similarity, ii)Longest Common Subsequence, iii)Dice Coefficient, iv)Fingerprint based Jaccard Similarity and v) Fingerprint based Containment Similarity. In addition, three machine learning approaches namely i) naïve Bayes, ii) Support Vector Machine, and iii) linear logistic regression classifiers are used for Arabic-English Cross-language plagiarism detection. Several experiments are conducted to evaluate the performance of the key phrases extraction methods. In addition, Several experiments to investigate the performance of machine learning techniques to find the best method for Arabic-English Cross-language plagiarism detection.

According to the experiments of Arabic-English Cross-language plagiarism detection, the highest result was obtained using SVM classifier with 92% f-measure. In addition, the highest results were obtained by all classifiers are achieved, when most of the monolingual plagiarism detection methods are used.

Keywords: Cross Language Plagiarism Detection, Mono-Language Plagiarism Detection, Classification, Machine Learning, Key Phrases, Candidate document.

1. INTRODUCTION

Cross-lingual plagiarism (CLP) happens when texts are written in one language to be translated into another language and used without acknowledging to the original sources. Extensive studies have been executed on monolingual plagiarism analysis which content searching for plagiarism in documents of the same language, but CLP still remains a challenge. Previous studies have addressed this problem using methods such as Statistical Machine Translation [1], cross-lingual showed semantic analysis (CL-ESA)

Comment [EO1]: Revert to original sentence for smooth flow of grammar.

[2], syntactic alignment using character N-grams (CL-CNG), dictionaries and thesaurus [3] [4], online machine translators [5, 6], and more recently, semantic networks and word embedding [7] [8]. and [9, 10]. Most of the suggested pattern are either limited to bilingual cross-lingual plagiarism detection tasks, when require parallel or comparable corpus which are usually not sufficient or available for low resource languages, while others trust on internet translation services, which are not existing for large scale cross-lingual plagiarism detection.

Different methods have been used to solve the cross lingual plagiarism detection. Based on the literature, it could be noticed that the majority of these methods can be classified into machine translation based approaches, parallel corpora based models and hybrid models. The main problems of the existing cross-language plagiarism detection techniques that uses machine translation as main method where the quality of the existing machine translation in translating big texts (whole documents) is very low and detecting plagiarism in translated documents is very challenging task because of the lexical and structural changes. In addition, when translated texts are replaced with their synonyms, using online machine translators to detect CLP would result in poor performance. To handle the limitation of these methods, this paper aim to design and implement a keyphrases based cross lingual plagiarism detection method. A significant feature of the proposed methodology is that it can be more efficient for detecting mono lingual paraphrased plagiarism where the sentence structure is changed and cross lingual translated plagiarism, as it keyphrases based detection method and keyphrases and their translation cannot be paraphrased.

This proposed research methodology consists of five phases, denoted as i) documents pre-processing phase, ii) Key phrase Extraction, Translation and Fingerprinting phase, iii) Retrieval of Candidate Documents phase, vi) Monolingual plagiarism detection phase and v) Machine Learning phase .The research methodology used in this study has facilitated to accomplish the objectives in terms of designing, developing, and implementing an efficient Arabic – English cross lingual plagiarism detection..

The remainder of this paper is structured as followings .Section 2 provides related work of cross-language Arabic – English techniques, as applied to words or sentences . Section 3 is proposed methodology , explaining the various proposed algorithms which are used for the pre-processing and framework CLPD ; the techniques mentioned in section 3, namely pre-processing is tokenization and stop word and NLP techniques in section 3.1 ; in section 3.2, the techniques are the key phrase extraction -based techniques, namely c- value algorithm and NC-value and key phrase ranking to find similarity score after that translate Arabic key phrases to English and retrieval candidate document and compare fingerprint for the key phrases in section 3.4 . Section 3.5 monolingual methods N-Grame and longest common subsequence to compare candidate document and suspicious document by hash table for fingerprint; and section 3.6 Machine Learning phase in this section is plagiarised text or not. in section 4 presents the experimental design, including the tools and packages used in this study, the datasets involving 318 document from the Arabic and English language benchmark dataset. Section 5 presents the results and discussion of findings and, finally, in section 6, conclusions and recommendations for future research will be provided.

Comment [EO2]: Replace with ‘as follows:’

Comment [EO3]: There should be no space between ‘s’ and ‘.’

Comment [EO4]: There should be no space between ‘CLPD’

Comment [EO5]: Replace with ‘documents’

Comment [EO6]: Replace with ‘are’

2. RELATED WORK

In this section, we give an overview of existing research in the area to focused on dataset of document. Specifically focusing on candidate document categorization. In [11], text pre-processing techniques, such as stopword removal, and shallow NLP techniques, such as stemming, are applied to documents before counting similarity. Short sentences are also deleted . The degrees of similarity between words are computed by their frequency of co-occurrence and relative distance, as mentioned by a word-correlation matrix generated using Wikipedia . A threshold is set to candidate sentences with a low similarity, and the degree of

Comment [EO7]: Replace with ‘ of focused on dataset of document.’

resemblance between two documents is visualized using Dot plot view. Although the results interpreted development over n-gram matching by decreasing the false positives, the approach is still limited to comparison between individual words.

Experiments were created on a domain-specie corpus compounding of English, Arabic, French, Spanish and Russian texts translated into Italian[12]. The experiment was executed using an SVM classifier, based on features such as lemmatised words and POS sequences. The best accuracy was achieved by using a combination of features that includes 1-Gram word with TF-IDF weighting, and 2-Grams and 3-Grams of POS tags. The experiment finished that the task biases on the distribution of n-Grams of function words and morpho-syntactic features.

Pouliquen introduced a statistical approach to map multilingual documents for a language-independent document representation, which measures similarity between monolingual and cross-lingual documents. A parallel corpus with multilingual interpreted texts was used, and pre-processing techniques including lemmatisation and stopword removal were applied. Parallel texts in various languages are determined by the *tf-idf* of the topic, and the top 100 words are chosen as descriptors. Each descriptor contains one-to-one interpretations into various languages and is stood for by a vector. The similarity score was computed by comparing the vectors between Spanish and English documents[13].

Aljohani, [14]and Mohd introduced the first Arabic-English cross-language plagiarism detection using the Winnowing Algorithm to discover the Arabic sentences translated from English sources without indecation of the original sources, as well as to diagnouseing its main containt and processes. The result clarifies that the Winnowing algorithm can be used effectively to discover the Arabic-English cross-language plagiarism with 81% recall, 97% precision and 89% F-measure.

Omar, Alkhatib [15] study a method for plagiarism detection algorithm in both Arabic and English languages. They proved a system to detect plagiarism in both Arabic and English languages using "Bing" search machain. The system which bases on plagiarism detection algorithm is effective and can supply both Arabic and English languages.

Kent, [16] study a method for a web-based system to discover cross-lingual plagiarism. The system decreases candidate document by summarizing. The Summary is interpreted to English, whereas similar web resources are discovered.

Gottschalk [17]and Demidova improved methods to join text passages written in various languages and consisting of overlapping data. The authors used Named entities and text interpretation to English as features to estimate the similarity between documents. These approaches use text translation as part of the process of obtaining a common comparison space. However, since text translation is a challenging task, it may arrive to high false rate.

Ferrero[9] suggested methods for cross-lingual plagiarism detection using word embeddings. These methods require training using decision tree or weights optimization, so here they are supervised methods.

[18] introduced a language autonomous model that measures the semantic similarity between text captures across multiple languages. The system uses a Support Vector Machine (SVM) to summarize a number of inter textual features, which contains features divided from embeddings trained using the word2vec model and a multi-lingual corpora, from lexical similarity measurements, from the internal representation (hidden layer) of a neural network trained using multi-lingual parallel corpora and from CL-ESA. This approach is however best appropriate for low resource languages.

Comment [EO8]: Replace with 'indication'

Comment [EO9]: Replace with 'to diagnosing its main content'

Comment [EO10]: Past tense

Comment [EO11]: Past tense

Comment [EO12]: Kindly reframe the sentences strictly according to the coorrection made by Feedback_AJRCOS_46873_v1_Ant_A.doc

Comment [EO13]: Authors names are needed.

Comment [EO14]: You should rather mind your language. Review of related work are normally reported in past tense.

3. METHODOLOGY

This research will study the problem of cross lingual plagiarism detection solution, and proposed solutions for this problem. the primary goal of the research is to design and implement methods for Arabic – English cross lingual plagiarism detection .

This research methodology consists of five main phases, denoted as i) Documents pre-processing phase, ii) Key phrase Extraction, Translation and Fingerprinting phase and iii) Retrieval of Candidate Documents phase, vi) Monolingual plagiarism detection phase and v) Machine Learning phase.

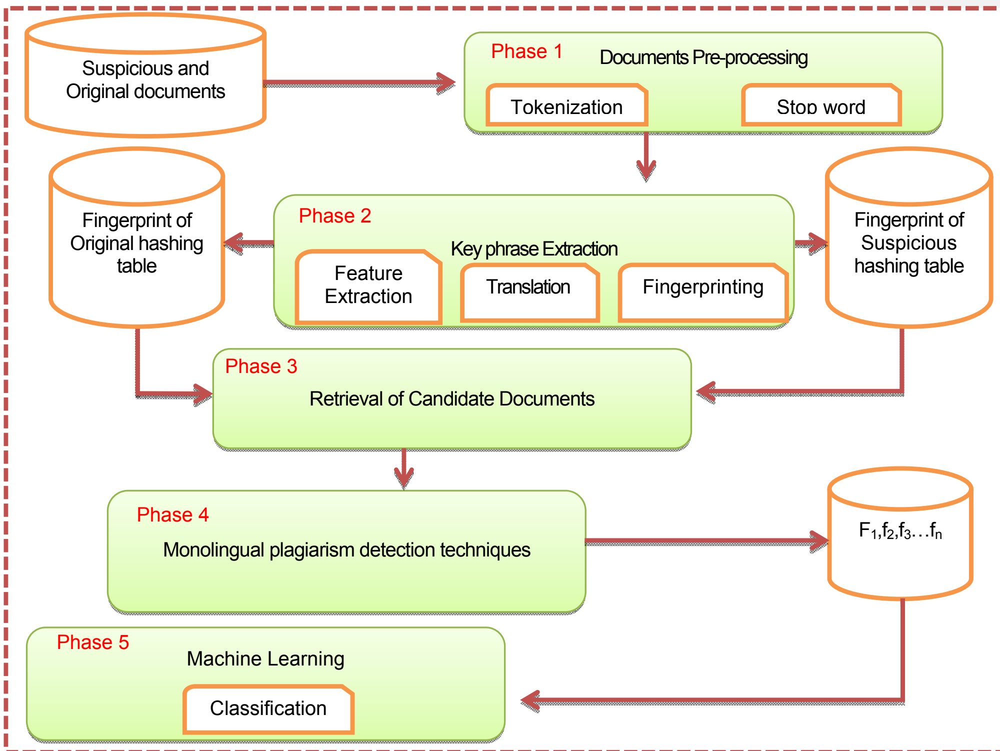

Fig. 1. **The Proposed Methodology Of Arabic – English Cross Language Plagiarism Detection**

3.1 Preprocessing.

In the pre-processing stage, various NLP pre-processing techniques are applied in a first step, each document is spilt into sentences. This work use (.), (;), (:), (!) And (?) Punctuation marks as a spilt point. After splitting documents into sentences, the sentences pre-processing consists of three steps: 1) tokenization, 2) normalization, 3) stop word removal. All sentences went through a pre-processing stage. In the normalization process, noisy characters are removed. Secondly, in this phase certain stop-words that occur commonly in all documents were removed to avoid plagiarism detection over fitting. After the pre-processing stage, each document is represented as a bag of sentences and each sentence in its turn is modelled as Bag Of Words..

<table>
<tr><th colspan="7">Tokenization</th></tr>
<tr><th colspan="7">Input</th></tr>
<tr><td colspan="7">قياس_الضوء_الطيفي
في الفيزياء , قياس الضوء الطيفي , هو دراسة كمية طيف كهرومغناطيس للطيف الكهرومغناطيسي.
وهو أكثر تخصصا من قياس الطيف الكهرومغناطيس , حيث يتعامل فقط مع طيف مرئي , وقريب أشعة فوق البنفسجية
وقريب أشعة تحت الحمراء.</td></tr>
<tr><th colspan="7">Out put \ Input Stop word</th></tr>
<tr><td colspan="2">،</td><td>الفيزياء</td><td>في</td><td>الطيفي</td><td>الضوء</td><td>قياس</td></tr>
<tr><td>كمية</td><td>دراسة</td><td>هو</td><td>،</td><td>الطيفي</td><td>الضوء</td><td>قياس</td></tr>
<tr><td colspan="2">أكثر</td><td>وهو</td><td>الكهرومغناطيسي</td><td>للطيف</td><td>كهرومغناطيس</td><td>طيف</td></tr>
<tr><td>حيث</td><td>،</td><td>الكهرومغناطيس</td><td>الطيف</td><td>قياس</td><td>من</td><td>تخصصاً</td></tr>
<tr><td>وقريب</td><td>،</td><td>مرئي</td><td>طيف</td><td>مع</td><td>فقط</td><td>يتعامل</td></tr>
<tr><td>الحمراء</td><td>تحت</td><td>اشعة</td><td>وقريب</td><td>البنفسجية</td><td>فوق</td><td>اشعة</td></tr>
<tr><th colspan="7">Stop word</th></tr>
<tr><th colspan="7">Out put</th></tr>
<tr><td colspan="2"></td><td>الفيزياء</td><td></td><td>الطيفي</td><td>الضوء</td><td>قياس</td></tr>
<tr><td></td><td>دراسة</td><td></td><td></td><td>الطيفي</td><td>الضوء</td><td>قياس</td></tr>
<tr><td colspan="2"></td><td></td><td>الكهرومغناطيسي</td><td>للطيف</td><td>كهرومغناطيس</td><td>طيف</td></tr>
<tr><td></td><td></td><td>الكهرومغناطيس</td><td>الطيف</td><td>قياس</td><td></td><td>تخصصاً</td></tr>
<tr><td></td><td></td><td>مرئي</td><td>طيف</td><td></td><td></td><td>يتعامل</td></tr>
<tr><td>الحمراء</td><td>تحت</td><td>اشعة</td><td></td><td>البنفسجية</td><td>فوق</td><td>اشعة</td></tr>
</table>

Fig. 2. Pre-processing tokenization and stop word of Arabic Document

3.2 Key phrases Extraction Phase

The main problems of the existing cross-language plagiarism detection techniques that uses machine translation as main method where the quality of the existing machine translation in translating big texts (whole documents) is very low and detecting plagiarism in translated documents is very challenging task because of the lexical and structural changes.

Key phrases are the important words/phrases that reflect the subject of the text. The Key phrases describe a document in a coherent and simple way for giving the prospective reader a way to quickly determine whether the document satisfied their information need. According to that, we index each document by Key phrases and only translate them, If the similarity score   is so high between the Key phrases of two documents, then one of these documents will be selected as suspicious document. However, the method used here for key phrases extraction consists of four steps   1) Features Extraction 2) Ranking 3) translation 4)fingerprinting.

Comment [EO15]: Delete ‘for’

Comment [EO16]: Replace with ‘satisfies’.

3.2.1 Features Extraction

The following features are used for ranking the candidate key phrase:

3.2.1.1 Phrase Frequency

Frequency is the number of occurrences of the candidate phrase. Frequency is normalized by the number of all candidate phrases in the document.as [19]

$$f_{\theta} = tf\,(kp) = \frac{\#(kp)}{\sum_{n\in\{all\_phrases\}} \#(n)} \quad (3.1)$$

3.2.1.2 C-value Approach

The C-Value method is a hybrid domain-independent method combining linguistic and statistical information (with emphasis on the statistical part) for the extraction of key phrases and nested phrases (i.e. phrases that appear within other longer phrases, and may or may not appear by themselves in the corpus). This method takes as input a corpus and produces a list of candidate key phrases, ordered by the likelihood of being valid terms, namely their C-Value measure... C-value is defined as [20]:

$$f_{\varphi} = c - value(c) = \log_2 |c| \left\langle f\,(c) - \frac{1}{p(T_C)} \sum_{P\in T_C} f\,(b) \right\rangle \quad (3.2)$$

Where C is a candidate key phrase, $|C|$ is the number of simple nouns that consist of C, $f(c)$ is its frequency of occurrence in the corpus, T_C is the set of extracted candidate terms that contain C, $P(T_C)$ and is the number of this candidate term.

3.2.1.3 NC-Value

The NC-Value is used to re-rank and improve the list of the extracted key phrases based on information from the term's neighbourhood. It, therefore, ranks the list of candidate key phrases, trying to bring higher key phrases that are more likely to contain key phrases. The NC-value measure is computed as [19, 21]:

$$NC - value\,(a) = 0.8C - value\,(a) + 0.2 \sum_{b\in a} f_a(b)\; weight\,(b) \quad (3.3)$$

<u>3.2.2 Key phrases Ranking And Filtering</u>

This main purpose of this phase is to extract the most important Key phrases. To rank each key phrase from the candidate Key phrases .

<u>3.2.3 Translation And Language Normalization</u>

In order to overcome the language barrier, all original documents (represented by extracted key phrases) are translated into one language in this case the Arabic key phrases translated in to English language has been chosen as it has bilingual translation between it and most of languages. For this purpose, the present work adopted Google Translate (GT) as it offers API access and is considered the state-of-the-art machine translation system used today.

<u>3.2.4 Fingerprinting</u>

Comment [EO17]: Kindly reframe the sentences strictly according to the coorrection made by Feedback_AJRCOS_46873_v1_Ant_A.doc

Document fingerprinting is the process of representing a document as a set of integers resulting from hashing substrings of the document. The comparison is then performed on the fingerprint rather than the whole text. In this work, the process of creating a fingerprint is as follow:

- Key phrasing: key phrases are extracted and each sentence is represented as a Bag Of Words.
- Hashing: a hash function is applied to the extracted key phrases to map them to a vector of integers.

3.3 Retrieval Of The Candidate Documents Phase

The process of candidate documents retrieval is through measuring similarities between the input document and the candidate documents at sentence level. In the fingerprinting method, the amount of similar fingerprints is used as similarity indicator between sentences; measuring similarity between two sentences or subdocuments is calculated by comparing the similarity percentage between a sentence's fingerprint and another sentence's fingerprint. For two sentences A and B, let $h(A)$ $h(A)$ and $h(B)$ be their fingerprints with the corresponding length |h (A)| and |h (B)|. A similarity between A and B based on $h(A)$ $h(A)$ and $h(B)$ calculate the percentage of the similar fingerprints as [22, 23]:

$$sim(A,B) = \frac{|h(A) \cap h(B)|}{|h(A)|} \qquad (3.4)$$

If $Sim(A,B)$ is greater than a threshold , subdocument B is selected as candidate subdocument.

3.4 Monolingual Plagiarism Detection Techniques

The output of these methods will be used as feature vector that is used to training a machine learning classifier. In this work, several monolingual plagiarism detection techniques have been adopted:

3.4.1 N-Grams Similarity

The number of overlapping n-grams between two documents, d^S the suspicious document and d_i^C document i from the candidate document, will be counted . the overlapping total is divided by the length of the suspicious subdocument and length of the candidate subdocuments respectively in order to calculate recall and precision.

N-gram similarity score is expressed as[23]:

$$Score(d^s, d_i^c) = \frac{2*(R-N)*(P-N)}{(R-N)+(P-N)} \qquad (3.5)$$

3.4.2 Longest Common Subsequence (LCS)

Given two documents, LCS is the longest string of matched tokens between these documents. LCS is that unlike n-grams (excluding unigram), LCS allows skip of matched n-grams. LCS score can be expressed as follows[24]:

$$ScoreLCS\,(d^{s}, d_{i}^{c}) = \frac{2*(R-LCS)*(P-LCS)}{(R-LCS)+(P-LCS)} \quad (3.6)$$

3.4.3 Dice Coefficient

The Dice similarity between two subdocuments A and B is defined as in[25]:

$$Dice(A,B) = \frac{2a}{2a+b+c} \quad (3.7)$$

Where (a) refers to the matched key phrases or fingerprints present in both A and B, (b) refers to the key phrases or fingerprints present only in A , and (c) refers to those present only in B.

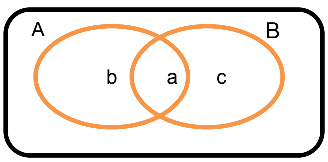

Fig. 3. Dice Coefficient Similarity

3.4.4 Fingerprint based Jaccard Similarity

Jaccard similarity is a very common set similarity measure that is used in a wide variety of applications. It is defined as[26]:

$$jaccard\,(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.8)$$

Where A is the suspect fingerprints and B is the source fingerprints.

3.4.5 Fingerprint based Containment Similarity

Containment similarity is nearly identical to jaccard similarity, except the denominator is only the number of elements in the suspect fingerprint. Again, let A be the suspect fingerprints and B be the source fingerprints. Due to the size difference in of these fingerprints sets, an asymmetric similarity measure is conducted based on containment similarity as [27]:

$$Containment\,(A,B) = \frac{|A \cap B|}{|A|} \quad (3.9)$$

3.5 Machine Learning Phase

The main idea is to feed the output of Monolingual plagiarism detection techniques to a machine learning classification framework. As shown in the previous sections, the monolingual plagiarism detection measures are only measure the similarity between suspicious document and candidate documents. However, their scores cannot indicate explicitly whether spacious document is plagiarized or not. To indicate explicitly whether suspicious document is plagiarized or not, we evaluated several classification methods for plagiarism detection.

3.5.1 Linear Logistic Regression

Logistic regression predicts the probability of an outcome that can only have binary response, also can handle several predictors (numerical and categorical). The multiple logistic regression model has the form as [24] :

$$\log(displag) = b_0 + b_1 X_1 + + b_k x_k \qquad (3.10)$$

$$f(x) = p(displagirasized) = \frac{\exp^{b_0 + b_1 x_1 + ...+ b_k x_k}}{1 + \exp^{b_0 + b_1 x_1 + ...+ b_k x_k}} \qquad (3.11)$$

3.5.2 Naive Bayes

The major advantage of NB algorithms is that they are easy to implement, often they have a superior performance. Naive Bayes (NB) can be defined as the conditional probability of plagiarized class pc given monolingual feature vector mf constructed as follows as[28]:

$$P(pc \mid mf) = p(c \mid s_1, ..., s_j) = p(pc) \prod_j p(s_j \mid pc) \qquad (3.12)$$

Thus, the maximum posterior classifier is given as follows:

$$c^* = \arg\max_{c=C} \; p(c) \prod_{i=1}^{n} p(t_i \mid c) \qquad (3.13)$$

3.5.3 Linear Discriminate Analysis

The basic idea of LDA is to find a one-dimensional projection defined by a vector v that maximizes class separation. This method maximizes the ratio of between-class variance S_B to the within-class variance S_W in any particular data set thereby guaranteeing maximal separability as[29].

$$\max_v \frac{v^t S_B v}{v^t S_w v} \quad (3.14)$$

3.5.4 Support Vector Machines

SVM is a featured machine learning technique that is developed for the binary classification task. SVM proposed to solve two-class problems by finding the optimal separating hyper-plane between two classes of data. Suppose that X is set of labelled training points (feature vector $(x_1, y_1), ...(x_n, y_n))$ where each training point $x_i \in$ RN is given a label $y_i \in$ {−1, +1},where i = 1,. . .,n. The goal in SVM is to estimate a function $f(x) = w.x_i + b$ and to find a classifier $y(x) = sign(f(x))$ which can be solved through the following convex optimization as[18] :

$$\min_{w,b} \sum_{i=1}^{n} [1 - y_i (w . x_i + b)] + \frac{\lambda}{2} \|w\| \quad (3.15)$$

with λ as a regularization parameter.

4. EXPERIMENTAL RESULTS

In this section, several experiments have been conducted in order to evaluate the proposed approaches. First, several experiments have been conducted to evaluate key phrases extraction methods. Secondly, Several experiments to empirically compare several monolingual plagiarism detection methods and three classification approaches which are i)Linear Logistic Regression, ii) naïve Bayes , iii) SVM classifiers for Arabic-English Cross-language plagiarism detection. This research uses the same data set used by ALAA et al 2017 [24] for Arabic-English Cross-language plagiarism detection system. The data consists of 318 Arabic files are used for both training and test. All English files were used for the comparison of both training and testing stages.

Table 4.1 : Detailed description of the experiment dataset

Dataset	Training	Test	Total
Arabic Files	200	118	318
English Files	34	20	54

4.1 Experimental Results Of SVM Classifier

In this experiment, SVM classifier is applied on testing set using 10-fold cross-validation. In this work, we used all monolingual plagiarism detection methods namely N-Grams Similarity (M1), Longest Common Subsequence (LCS) (M2), Dice Coefficient (M3), Fingerprint based Jaccard Similarity (M4), Fingerprint based Containment Similarity(M5) as a features for SVM.

Table 4.2 shows the performance in terms of the precision, recall, F-measure of Arabic-English Cross-language plagiarism detection by applying the SVM classifier with using different combination set of features. The highest result yield by SVM classifier trained is 92% f-measure. As shown in Table 4.2, low performances are obtained when SVM uses only one or two monolingual methods as features and high performances are obtained when SVM uses more than three monolingual methods as features. This means that using all monolingual plagiarism detection methods has an obvious positive effect on the quality detection method.

Table 4.2 the performance of SVM Arabic-English Cross-language plagiarism Detection

M1	M2	M3	M4	M5	PRECISION	F-MEASURE
0	1	0	1	0	0.74	0.85
0	1	0	0	1	0.69	0.82
0	1	0	0	0	0.59	0.74
0	1	0	1	0	0.75	0.86
1	1	0	1	0	0.73	0.84
0	1	0	1	1	0.67	0.8
1	1	0	0	0	0.4	0.57
1	1	0	0	1	0.76	0.86
0	1	0	0	1	0.71	0.83
1	0	0	0	1	0.61	0.76
1	0	0	1	0	0.73	0.84
1	0	1	0	0	0.79	0.88
0	1	1	0	0	0.74	0.85
1	1	1	1	1	0.84	0.91
0	1	1	1	1	0.85	0.92

4.2 Experimental Results Of NB Classifier

In this experiment, NB classifier is applied on testing set using 10-fold cross-validation. The idea is to show the best results obtained when the NB classifier is applied. In this work, we used all monolingual plagiarism detection methods namely N-Grams Similarity (M1), Longest Common Subsequence (LCS) (M2), Dice Coefficient (M3), Fingerprint based Jaccard Similarity (M4), Fingerprint based Containment Similarity(M5) as a features for NB.

Table 4.3 shows the performance in terms of the precision, recall, F-measure of Arabic-English Cross-language plagiarism detection by applying the NB classifier using different combination set of features. The highest result yield by NB classifier trained is 89% f-measure. This means that using all monolingual plagiarism detection methods has an obvious positive effect on the quality detection method. However, the results obtained by NB are lower than that of SVM.

Table 4.3 the performance of NB Arabic-English Cross-language plagiarism detection

M1	M2	M3	M4	M5	PRECISION	F-MEASURE
0	1	0	1	0	0.53	0.69
0	1	0	0	1	0.65	0.79
0	1	0	0	0	0.56	0.72
0	1	0	1	0	0.68	0.81
1	1	0	1	0	0.39	0.56
0	1	0	1	1	0.69	0.82
1	1	0	0	0	0.61	0.76
1	1	0	0	1	0.69	0.82
0	1	0	0	1	0.75	0.86
1	0	0	0	1	0.77	0.87
1	0	0	1	0	0.74	0.85
1	0	1	0	0	0.75	0.86
0	1	1	0	0	0.7	0.82
1	1	1	1	1	0.8	0.89
0	1	1	1	1	0.79	0.88

4.3 Experimental Results Of Linear Logistic Regression Classifier

In this experiment, linear logistic regression classifier is applied on testing set using 10-fold cross-validation. The idea is to show the best results obtained when the linear logistic regression classifier is applied. In this work, we used all monolingual plagiarism detection methods namely N-Grams Similarity (M1), Longest Common Subsequence (LCS) (M2), Dice Coefficient (M3), Fingerprint based Jaccard Similarity (M4), Fingerprint based Containment Similarity(M5) as a features for NB.

Table 4.4 shows the performance in terms of the precision, recall, F-measure of Arabic-English Cross-language plagiarism detection by applying the linear logistic regression classifier using different combination set of features. The highest result yield by linear logistic regression classifier trained is 86% f-measure. This means that using all monolingual plagiarism detection methods has an obvious positive effect on the quality detection method. However, the results obtained by linear logistic regression are lower than that of SVM and NB.

Table 4.4 The performance of linear logistic regression Arabic-English Cross-language plagiarism detection

M1	M2	M3	M4	M5	PRECISION	F-MEASURE
0	1	0	1	0	0.49	0.66
0	1	0	0	1	0.61	0.76
0	1	0	0	0	0.52	0.68
0	1	0	1	0	0.64	0.78
1	1	0	1	0	0.36	0.53
0	1	0	1	1	0.64	0.78
1	1	0	0	0	0.57	0.73
1	1	0	0	1	0.67	0.8
0	1	0	0	1	0.73	0.84
1	0	0	0	1	0.74	0.85
1	0	0	1	0	0.73	0.84
1	0	1	0	0	0.71	0.83
0	1	1	0	0	0.67	0.8
1	1	1	1	1	0.76	0.86
0	1	1	1	1	0.74	0.85

5. RESULTS DISCUSSION

This paper aim to examine the proposed model and observation of the experimental results that have been achieved.

In the result tables in the fields (M1, M2, M3, M4, M5) there are values:
1 : indicates that it was used in the experiment .
0 : indicates that it was not used in the experiment.

According to the experiments of Arabic-English Cross-language plagiarism detection with the SVM, NB, linear logistic regression classifiers, the highest result yield by SVM classifier with 92% f-measure.

According to the experiments of Arabic-English Cross-language plagiarism detection using SVM, NB, linear logistic regression classifiers with different combination of monolingual plagiarism detection methods namely N-Grams Similarity (M1), Longest Common Subsequence (LCS) (M2), Dice Coefficient (M3), Fingerprint based Jaccard Similarity (M4) and Fingerprint based Containment Similarity(M5) , the highest results obtained by all classifiers are achieved when most of the monolingual plagiarism detection methods used.

Furthermore, the obtained results with 92% f-measure were better than the previous work of Aljohani [14]et al. (2014) at 89% and of ALAA [24]et al (2017) with 90%

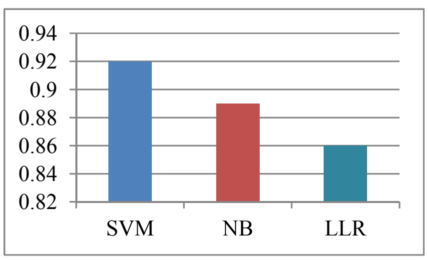

Fig. 4. Conclusion of SVM And NB,LLR Result

6. CONCLUSION

Due to rapid growth of research articles in various languages, cross-lingual plagiarism detection problem has received increasing interest in recent years. Cross-lingual plagiarism detection is more challenging task than monolingual plagiarism detection. This paper aims to design and implement a keyphrases based cross lingual plagiarism detection method. This paper empirically investigates five different monolingual plagiarism detection methods with three machine learning approaches namely naïve Bayes, SVM, and linear logistic regression classifiers are used for Arabic-English Cross-language plagiarism detection. Several experiments are conducted to evaluate the performance of the key phrases extraction methods. In addition, several experiments to investigate the performance of machine learning techniques to find the best method for Arabic-English Cross-language plagiarism detection. According to the experiments of Arabic-English Cross-language plagiarism detection, the highest result yield by decision SVM classifier with 92% f-measure. In addition, the highest results obtained by all classifiers are achieved when most of the monolingual plagiarism detection methods used.

Future work will aim to evaluate the current methodology with different language pairs. In addition, future work will studied multilingual plagiarism detection i.e. include more than two languages.

References

1. Barrón-Cedeño, A., P. Gupta, and P. Rosso, Methods for cross-language plagiarism detection. Knowledge-Based Systems, 2013. 50: p. 211-217.
2. Potthast, M., et al., Cross-language plagiarism detection. Language Resources and Evaluation, 2011. 45(1): p. 45-62.
3. Pataki, M., A new approach for searching translated plagiarism. 2012.
4. Gupta, P., A. Barrón-Cedeno, and P. Rosso. Cross-language high similarity search using a conceptual thesaurus. in International Conference of the Cross-Language Evaluation Forum for European Languages. 2012. Springer.
5. Ehsan, N., F.W. Tompa, and A. Shakery. Using a dictionary and n-gram alignment to improve fine-grained cross-language plagiarism detection. in Proceedings of the 2016 ACM Symposium on Document Engineering. 2016. ACM.
6. Ferrero, J., Agnes, F., Besacier, L. and Schwab, D.,, SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity. arXiv preprint arXiv:1702.03082, 2017c.
7. Franco-Salvador, M., P. Rosso, and M. Montes-y-Gómez, A systematic study of knowledge graph analysis for cross-language plagiarism detection. Information Processing & Management, 2016. 52(4): p. 550-570.
8. Speer, R.a.L.-D., J., ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. arXiv preprint arXiv, 2017. 1702.(03082): p. 1704-03560.
9. Ferrero, J., Agnes, F., Besacier, L. and Schwab, D., Using Word Embedding for CrossLanguage Plagiarism Detection. arXiv preprint arXiv:1702.03082, 2017a. . 1702.(03082.).
10. Glavaš, G., Franco-Salvador, M., Ponzetto, S.P. and Rosso, P., . , A Resource-Light Method for
Cross-Lingual Semantic Textual Similarity. Knowledge-Based Systems., 2017.
11. Pera, M.S. and Y.K. Ng, SpamED: A spam E□mail detection approach based on phrase similarity. Journal of the American Society for Information Science and Technology, 2009. 60(2): p. 393-409.
12. Baroni, a.B., M, S., A new approach to the study of translations: Machine learning the difference between original and translated text. Literary and Linguistic Computing, (2006). 21(3): p. 259–274.
13. Pouliquen, S., and Ignat., Automatic identification of document translations in large multilingual document collections. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'03), 2003. number 2002.(408): p. 401

14. Aljohani, A. and M. Mohd, Arabic-English Cross-language Plagiarism Detection using Winnowing Algorithm. Information Technology Journal, 2014. 13(14): p. 2349.
15. Omar, K., B. Alkhatib, and M. Dashash, The Implementation of Plagiarism Detection System in Health Sciences Publications in Arabic and English Languages. International Review on Computers & Software, 2013. 8(4).
16. Kent, C.K. and N. Salim. Web based cross language plagiarism detection. in Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on. 2010. IEEE.
17. Gottschalk, S. and E. Demidova, MultiWiki: interlingual text passage alignment in Wikipedia. ACM Transactions on the Web (TWEB), 2017. 11(1): p. 6.
18. España-Bonet, C. and A. Barrón-Cedeño. Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity. in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017.
19. Guan, J., A study of the use of keyword and keyphrase extraction techniques for answering biomedical questions. 2016.
20. Lossio-Ventura, J.A., et al. Combining c-value and keyword extraction methods for biomedical terms extraction. in LBM: Languages in Biology and Medicine. 2013.
21. Frantzi, K., S. Ananiadou, and H. Mima, Automatic recognition of multi-word terms:. the c-value/nc-value method. International journal on digital libraries, 2000. 3(2): p. 115-130.
22. Lane, P.C., C. Lyon, and James A. Malcolm., "Demonstration of the Ferret plagiarism detector." Proceedings of the 2nd International Plagiarism Conference. , 2006..
23. Hoad, T.C. and J. Zobel, Methods for identifying versioned and plagiarized documents. Journal of the American society for information science and technology, 2003. 54(3): p. 203-215.
24. Alaa, Z., S. Tiun, and M. Abdulameer, CROSS-LANGUAGE PLAGIARISM OF ARABIC-ENGLISH DOCUMENTS USING LINEAR LOGISTIC REGRESSION. Journal of Theoretical & Applied Information Technology, 2016. 83(1).
25. Lin, D. An information-theoretic definition of similarity. in Icml. 1998. Citeseer.
26. Yih, W.-T. and C. Meek. Improving similarity measures for short segments of text. in AAAI. 2007.
27. Yang, Y., et al., Gb-kmv: An augmented kmv sketch for approximate containment similarity search. arXiv preprint arXiv:1809.00458, 2018.
28. Ngu, A.H., et al., Smartwatch-based iot fall detection application. Open Journal of Internet Of Things (OJIOT), 2018. 4(1): p. 87-98.
29. Altman, E.I., G. Marco, and F. Varetto, Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). Journal of banking & finance, 1994. 18(3): p. 505-529.