DOI: 10.15849/IJASCA.240330.03

Int. J. Advance Soft Compu. Appl, Vol. 16, No. 1, March 2024
Print ISSN: 2710-1274, Online ISSN: 2074-8523
Copyright © Al-Zaytoonah University of Jordan (ZUJ)

Plagiarism Detection System by Semantic and Syntactic Analysis Based on Latent Dirichlet Allocation Algorithm

Khalid M.O. Nahar[1,8], Ma'moun Alshtaiwi[2], Enas Alikhashashneh[3], Nahlah Shatnawi[4], Moy'awiah A. Al-Shannaq[5], Mohammed Abual-Rub[6], Basel Bani-Ismail[7]

[1,4,5]Computer Science Department, Faculty of Information Technology and Computer Sciences, Yarmouk University, Jordan
e-mail: khalids@yu.edu.jo, nahlah.s@yu.edu.jo, moyawiah.s@yu.edu.jo
[2]Department of Modern Languages, Faculty of Arts, Yarmouk University, Jordan
e-mail: m.alshtaiwi@yu.edu.jo
[3]Information Systems Department, Faculty of Information Technology and Computer Sciences Yarmouk University, Jordan
e-mail: enas.a@yu.edu.jo
[6]Department of Management Information Systems, School of Business, King Faisal University, Hofuf, Saudi Arabia
e-mail: mabualrub@kfu.edu.sa
[7]Department of Computer Science and MIS, Oman College of Management and Technology, Muscat, Oman
e-mail: basel.baniismail@omancollege.edu.om
[8]Faculty of Computer Studies, Arab Open University, P.O. Box 84901, Riyadh 11681, Saudi Arabia
e-mail: k.nahar@arabou.edu.sa

Abstract

The process of plagiarism detection is one of the challenges in revealing the originality of a document, especially in the fields of science and research. Natural language processing methods can recognize and determine the level of similarity between different documents. In this paper, we tackle the task of extrinsic plagiarism detection based on semantic and syntactic approaches. The objective is to identify segments of a document that show strong similarity with a group of reference documents dealing with the same topic. In this paper, we present our hybrid approach that implements semantic and syntactic features based on Latent Dirichlet Allocation (LDA) and Wu & Plamer algorithm. The proposed approach has been evaluated on a PAN13 public dataset with a total accuracy of 85%.

Keywords: *Plagiarism detection, semantics, WordNet, LDA, Wu & Plamer's Algorithm.*

1 Introduction

Scientific research is an integral part of the rapid progress of any scientific field. It aims to generate information and knowledge from existing data to solve problems, improve an

Received 21 October 2023; Accepted 22 January 2024

existing solution or prove a hypothesis. Many universities require students to publish their research online to obtain an advanced degree. This research must be innovative and original.

Technological advances have enabled the emergence of digital publishing as an essential alternative to paper-based publishing. It reduces the cost of publishing, spreads more quickly around the world, and makes information available to the public at any time. There is a massive amount of academic research on the Web in the form of journal articles, books, conference proceedings, and reports. This enormous growth is due to the concept of "publish or perish", which requires researchers to publish their work.

Principles of ethics and adherence to rules of conduct are essential in research. Respect for intellectual property and honesty may be considered the most important ethics in research. Plagiarism in academic research is considered a violation of ethical rules. Plagiarism is claiming ownership of an idea, process, result, or words by stealing them from another researcher or simply copying someone else's work without reference [16].

To protect the innovation of ideas, plagiarism has serious consequences, such as the refusal of journal editors to evaluate and publish other research, the questioning of the institute in which the author works so that sanctions are taken against him or her, such as expulsion from the university and sometimes legal proceedings [26].

Plagiarism can take many forms, from the simplest, such as copying and pasting the same words without changing them, to the most complicated, such as paraphrasing sentences. Detecting plagiarized text is becoming a complicated problem due to advances in technology. Many software tools can provide synonyms for paraphrased terms and phrases, making it difficult for the reviewer to detect plagiarism. In addition, the sheer volume of searches makes it impossible for reviewers to manually compare them with a proposed search. The process is similar to finding a needle in a haystack.

Many software tools aim to explore and detect plagiarized texts using different methods. Current trends include fingerprinting, string matching, bag-of-words, citation analysis, and stylometry. Plagiarism detection methods can assess similarity locally or globally. The local similarity assessment approach compares the similarity between limited segments of the suspect text and the candidate text, such as the fingerprint. On the other hand, global similarity evaluation approaches explore the entire feature set of the document sections [15] [35].

In this research, we propose a new extrinsic plagiarism detection method based on machine learning classification methods combined with paraphrase detection techniques. We eliminate the copy-and-paste plagiarism problem and some of the intelligent plagiarism problems.

The rest of the paper is structured as follows. Section 2 reviews some previously related work. The proposed model architecture is described in Section 3. Section 4 discusses the results of this research. Finally, Section 5 concludes the paper and presents future research.

2 Related Work

The process of plagiarism can be very simple, such as copying and pasting, or it can be complicated by applying some intelligence, such as paraphrasing sentences, summarizing text, self-plagiarism, plagiarizing ideas, or translating an article from one language to another. Figure 1 shows the types of text plagiarism.

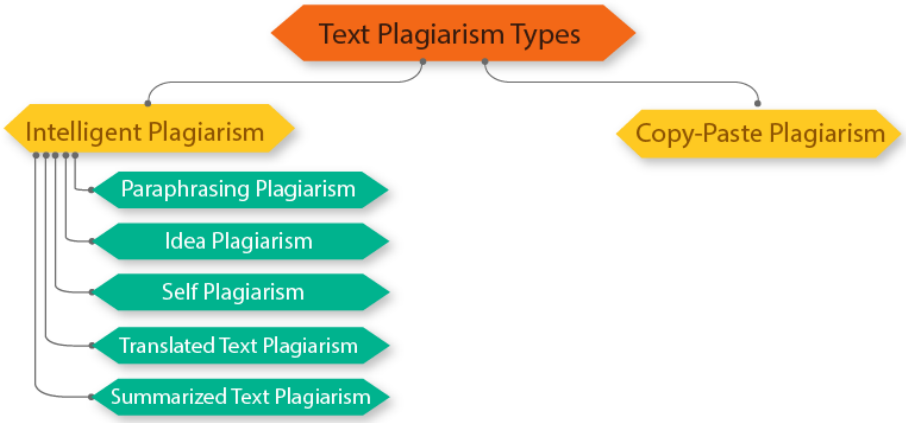

Figure 1: Types of text plagiarism.

The high need for plagiarism recognition has led to the invention of many plagiarism detection methods, based on the availability of new material comparable to the suspect document. Plagiarism detection methods are classified into two different categories, extrinsic and intrinsic, as shown in Figure 2.

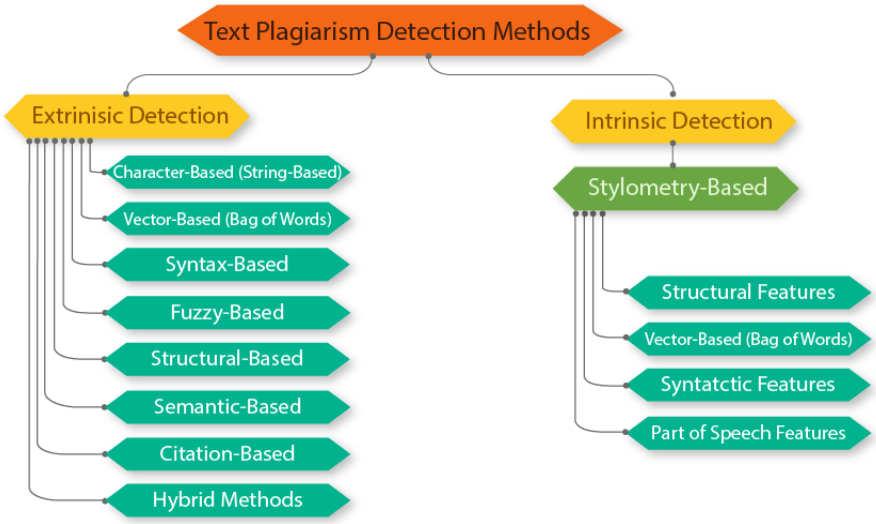

Figure 2: Types of text detection methods.

2.1 Extrinsic Detection

The extrinsic detection approach depends on an external comparison between the suspect document and the candidate group of new documents. The system depends on the existence of a certain corpus related to the document domain. The search for similar documents must be performed efficiently by reducing the search space [17].

2.1.1 Character-Based Method

This type of plagiarism searches for similarities between documents by performing an exact or partial word match. In the case of an exact match, the entire word must be identical in both texts, while in the case of a partial match, it is sufficient to have a match in a part of the word. N-grams are generally used in this method, as well as substrate-matching techniques.

Naik et al. [20] developed an approach to detect plagiarized copy-pasted texts using an N-gram language model for the Marathi language. The authors used the approach to test the Marathi corpus and obtained an accuracy of 90%.

2.1.2 Vector-Based Method

The vector method represents documents as tokens by extracting lexical and syntactic features to facilitate the comparison procedure. This method provides high recall values, and the similarity can be calculated using vector similarity measures such as Jaccard and Cosine.

To improve the performance of plagiarism checkers, Sornsoontorn et al. [25] categorized the documents in the database by applying a heuristic function to define a label for each document. This classification method makes it easier to find related candidates to compare with the suspect document using the vector method. The dataset and tests, which consist of theses and journals from Kasetsart University between (1998-2010), were conducted for the Thai language. The accuracy obtained is 92.7%.

Duarte et al. [9] employed a heuristic method, namely Minmax Circular Sector Arcs (MinmaxCSA), combined with target document indexing algorithms that can reduce the document search space and the costs of unnecessary comparisons. The corpus used was PAN-PC-11. The retrieval speed of MinmaxCSA performed better than the Minwise hashing method with a speed gain of 1.32.

2.1.3 Syntax-Based Method

In this method, Part-of-Speech (PoS) or grammatical marking is essential. English sentences will be split into nouns, verbs, pronouns, adjectives, adverbs, prepositions, conjunctions, and interjections. After labeling each word, similarity measures will be applied to the candidate documents. This approach is good when the text is being translated from one language to another. In PoS tagging, there are some difficulties such as some words can be tagged more than one part, especially these words can be nouns or verbs in English and French, for example, "Aller" "to Go" could be noun "Aller" "One-way ticket".

Vani and Gupta [27] built a document-level text plagiarism detector using shallow PoS techniques. The classification method depends on the selection of appropriate features in two phases which led to improvements using machine learning techniques such as Naive Bayes (NB), Support Vector Machine (SVM) and Decision Tree. They used two corpora, some instances of the PAN corpus and the Plagiarized Short Answers (PSA) corpus. They obtained an accuracy of 97.89% and an F-measure of 0.979.

2.1.4 Fuzzy-Based Method

The fuzzy-based method uses fuzzy numbers ranging from 0 to 1 for words of similar meaning in the document. For each document, the similarity is calculated based on the generated fuzzy numbers [10].

Rakian et al. [23] suggested a plagiarism detection for Persian language called Persian Fuzzy Plagiarism Detection (PFPD). It relies on sentence-level comparison by providing fuzzy numbers for similar sentences between 0 and 1 with a threshold value of 0.65. The obtained results outperform other detection methods, with precision, recall, and F-measure of 22.41, 17.61, and 18.54 respectively.

2.1.5 Structural-Based Method

To find plagiarized text, the structure-based method depends on contextual similarity such as the distribution of words in the entire document. This method is not widely used and usually depends on the representation of tree structure features.
Chow and Rahman [8] proposed a way to retrieve candidate documents and compare suspect documents based on their structure. They modeled the document by a rich hierarchical tree representation based on features such as documents, pages, and paragraphs. They used the multilayer self-organizing map (MLSOM) algorithm for efficient representation.

2.1.6 Semantic-Based Method

The semantic-based method has come to the forefront with advances in natural language techniques, such as the emergence of WordNet, which offers synonyms for the same word. This method depends on the meaning of the sentence, even if it is written using another method. Two sentences can have the same meaning, like converting a sentence from active to passive and vice versa [6].
Fan et al. [12] proposed a semantics-based method using deep neural networks called globalization semantic matching neural network (GSMNN). The goal was to find paraphrased sentences using the length of the utterance as a parameter. This process was found to be better than state-of-the-art paraphrase detection. They obtained an accuracy of 78% and the F1 measure of 71.95%.
Gang et al. [13] used WordNet to find paraphrased sentences using a methodology called Cross-language Plagiarism Detection (CLPD). It detects monolingual and cross-linguistic plagiarism based on WordNet synonyms for different languages, such as English, French, German, and Spanish. The methodology achieved a recall of 0.78 and a precision of 0.87.
Recently, Kaur et al. [31] proposed a novel plagiarism detection system based on semantic features to identify instances of plagiarism. To calculate the degree of similarity using semantic features, the system creates a dynamic relation matrix for each pair of suspicious and source sentences. This work presents a novel similarity measure for plagiarism detection together with two Weighted Inverse Distance and GlossDice techniques. On the PAN-PC-11 dataset, the proposed system fared better than the other baseline systems in terms of accuracy (0.9459), recall (0.8861), f-measure (0.8917), and plagdet (0.8857). The system demonstrates precision (0.9257), recall (0.9055), f-measure (0.8931), and plagdet (0.8806) for PAN-14 text alignment.

2.1.7 Citation-Based Method

A citation-based method is a new approach that studies citation patterns in a scientific document and the distinction between different citations. It compares the similarity of these citations with published articles to help recognize paraphrased and translated texts [14].
Meuschke et al. [19] proposed a prototype for plagiarism detection in scientific papers using the citation method. It is considered a language-independent model that could detect translation plagiarism using three algorithms. The model was trained and tested using articles from the PubMed Central Open Access Subset.
Being motivated by the idea of citation-based plagiarism recognition, Soleman and Fujii [24] developed a method based on citation networks and traditional citation detection methods. This method aimed to detect plagiarism even if the source document does not exist in the reference materials. The precision, accuracy, and F1 measure for the inspection task were 0.71, 0.83, and 0.77 respectively.

2.1.8 Hybrid Methods

Hybrid methods show improvement in solving many problems in different fields [34], Likewise, in plagiarism detection, Hybrid methods can use one or more types of the above methods to improve the results of the search and comparison of candidate documents.
Abdi et al. [1] designed a plagiarism detection system based on word meaning. They built an external plagiarism detection system that depends on the semantic role labeling (SRL) technique, and semantic and syntax-based information. The system can detect plagiarism by copying and pasting, paraphrasing a text, and changing the structure of a word, for example by converting it from passive to active and vice versa. They employed the PAN-PC-11 corpus for testing. The precision and recall values were 0.913 and 0.652 respectively.
Vani and Gupta [28] presented a system that depends on syntactic and semantic natural language processing techniques. The system is composed of three main natural language processing techniques, part-of-speech (PoS) tagging, chunking, and semantic role labeling (SRL). They used the PAN corpus between 2009-2014 for training and testing. The results are better than other techniques with recall and precision of 0.9643 and 0.8547 respectively.
To minimize the cost of searching for likely source documents for plagiarism, Kong et al. [18] proposed a model based on a ranking framework and a Ranking Logistical Regression model. They tested the model on PAN 2013 and the PAN 2014 Source Search Corpus, and the results showed improvements. The recall and precision values were 0.7475 and 0.7410 respectively. Al-Jibory et al. [32] proposed a system that combined machine learning (ML) and natural language processing (NLP) methods with an external plagiarism detection approach based on similarity analysis and text mining. Their method achieves 0.96 accuracy, 0.86 recall, 0.86 F-measure, and 0.86 PlagDet score. Arabi et al. [33] proposed a hybrid system of two methods to identify Extrinsic plagiarism, WordNet Ontology and FastText's pre-trained word embedding network are utilized in these two methods to generate the semantic matrix, while the TF-IDF weighting method is employed to form the structural matrix.

2.2 Intrinsic Detection

Contrasting with the extrinsic detection approach, the intrinsic detection method does not need external documents to compare. It relies on the author himself by studying the differences in writing style in different passages of the suspect document. In some cases, the document can be compared to previous works by the same author to ensure that there is no change in style. This method is effective when there is a lack of external resources to compare with Hourrane and Benlahmar [17].
Verifying authorship, detecting plagiarism, and attributing authors in multi-author documents was the aim of Aldebei et al. [2]. They used the hidden style for each author, based on the author's previous articles, using a hidden Markov model (HMM) using the Baum-Welch algorithm. The model generated many probabilities for the style, so they used the Viterbi algorithm to reach the shortest and best path. They achieved an accuracy of 96%.
By extracting stylometric features from multi-author documents, Elamine et al. [11] were able to detect the parts of each author's writing as well as any differences in writing style due to another author. They used lexical features such as average sentence length and syntactic features such as PoS tags. The model was trained and tested on PAN16 and PAN17 with precision and recall for each data set of (0.748, 0.635) and (0.701, 0.6).
To define an author's style, Vysotska et al. [29] introduced a plagiarism detection method based on NLP techniques and statistical linguistic analysis. The proposed method targets the Ukrainian language with the use of a Ukrainian corpus of academic articles.

Vysotska et al. [30] focused on recognizing an author's stylometry by using lingvometry methods to identify the author's percentage of work in a multi-author article. The proposed method evaluates the linguistic units in the author's work by calculating lingvometry coefficients from the author's previous articles.

Style detection approach based on the analysis of writing style using N-grams was experimentally utilized by Bensalem et al. [5]. The goal of their study was to evaluate the performance of character N-grams in terms of their frequency and length. The method was applied on five PAN corpora for English and Arabic documents, namely PAN-PC-09, PAN-PC-10, PAN-PC-11, InAra-Training, and InAra-Test. There was an increase in the F-measure from 0.31 to 0.35 on PAN-PC-09 for example. A summary of the literature review is shown in Table 1.

Table 1: Literature summary

Reference	Method	Corpus	Test Results	Language	Drawbacks
[25]	Vector-based	Journals from Kasetsart University	Accuracy = 0.927	Thai	NA
[27]	Syntax-based	PAN and Plagiarized Short Answers (PSA)	F1-measure = 0.979	English	NA
[12]	Semantic-based	NA	F1-measure = 0.7195	English	NA
[13]	Semantic-based	NA	F1-measure = 0.82	English, German, and Spanish	The need for clear references and some misleading results
[24]	Citation-based	NA	F1-measure = 0.77	English	The need for clear references
[1]	Hybrid	PAN-PC-11	F1-measure = 0.76	English	NA
[28]	Hybrid	PAN corpus from 2009 to 2014	F1-measure = 0.91	English	NA
[18]	Hybrid	PAN 2013 and the PAN 2014	F1-measure = 0.74	English	NA
[31]	Semantic-based	PAN-PC-11 and PAN-14	F1-measure =0.8917	English	NA
[32]	Hybrid	PAN-PC-11	F1-measure =0.86	English	NA
[33]	Hybrid	PAN-PC-11	Precision =95.1% and 93.8%	English	NA

3 The Architecture of the Proposed Model

3.1 Corpus

In terms of training and testing our model, we utilize the PAN 2013 corpus that was used in the 5th International Competition on Plagiarism Detection [21]. PAN annually generates different types of corpora for intrinsic and extrinsic types of plagiarism since 2009. PAN13 is a specialized type of extrinsic corpus called text alignment which is composed of two parts, PAN13 for training and PAN13 for testing.

PAN13 consists of 3169 source documents and 1826 suspect documents of which about half are plagiarized in different percentages and procedures as shown in Figure 3. The source documents of 145 topics were generated from texts composed by 27 editors and the suspect documents were generated using a program and some editors to modify the semantics as well as to summarize some passages from the source documents.

Some of the suspect files are randomly obfuscated, while others have undergone obfuscated translation and summary. In our model, we used 100 source documents and 25 suspect documents for training and testing respectively. The obfuscated documents exhibit both cut-and-paste plagiarism and paraphrastic plagiarism. We divided the dataset into two parts, 75% for training and 25% for testing.

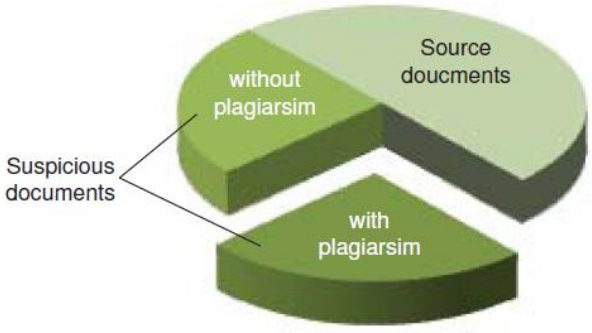

Figure 3: Distribution of suspicious documents [22].

3.2 General Architecture

Figure 4 shows our proposed multi-level model that can handle extrinsic plagiarism. The model can detect certain types of plagiarism, e.g., cut-and-paste plagiarism and paraphrastic plagiarism. A set of preprocessed source documents are grouped into clusters according to the topic distributions generated by the Latent Dirichlet Allocation (LDA). The preprocessed suspect input document will be checked for plagiarism by comparing it with the most relevant cluster (topic). Finally, the document will be marked as plagiarized or original.

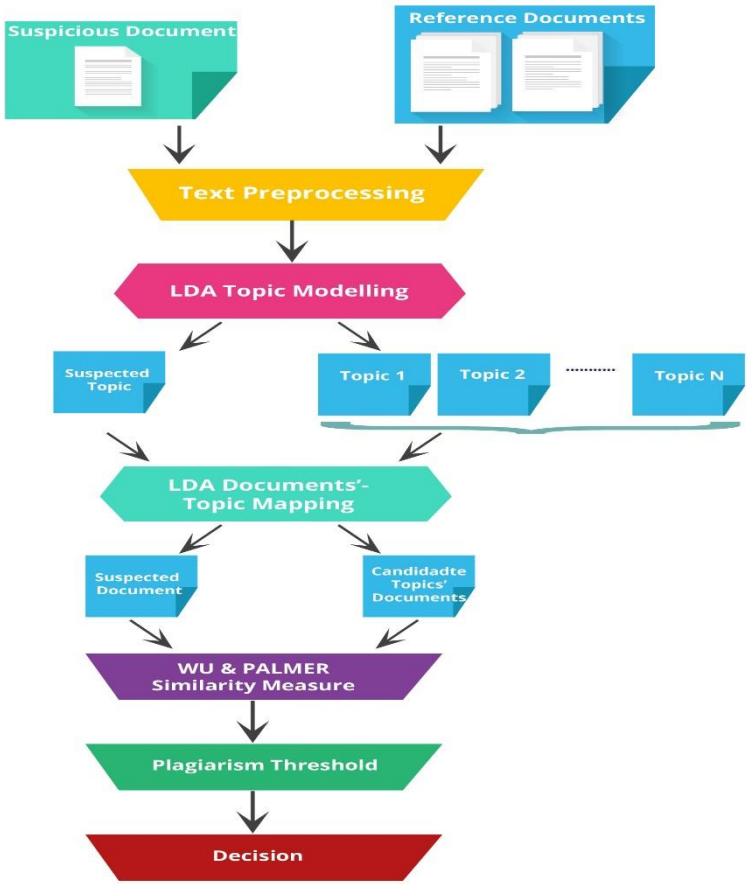

Figure 4: Plagiarism detection model architecture

3.3 Plagiarism Detection Stages

The presented plagiarism detection method is performed in four stages, namely, text preprocessing, LDA topic modeling, LDA document and topic matching, and plagiarism estimation.

3.3.1 Stage 1: Text Preprocessing

The first stage begins with the arrival of the suspect document and reference documents, which must be preprocessed to generate a bag of words as shown in Figure 5. The documents are converted to lowercase, modified by removing stop words and spaces, filtered of numbers, uprooted, lemmatized, modified by removing punctuation marks, tokenized, and converted to singulars. It is worth noting that the way of removing stop words in this stage is similar to the method proposed by Al-Shamery and Gheni [3].

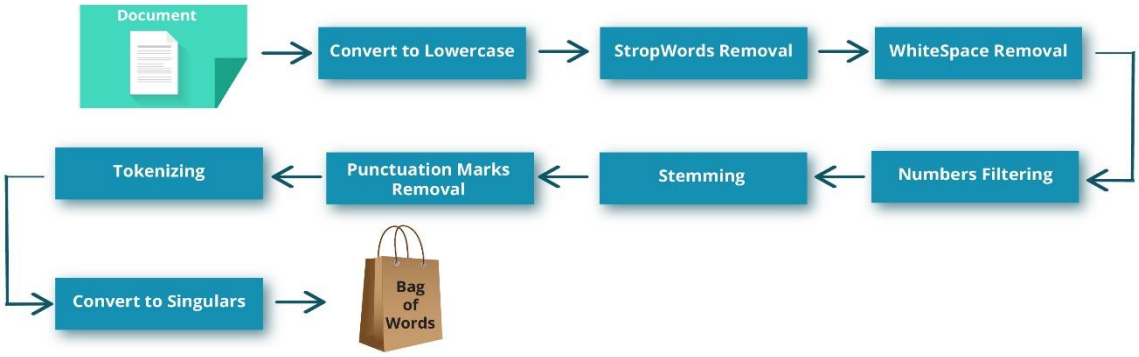

Figure 5: Preprocessing of documents

3.3.2 Stage 2: LDA Topic Modeling

To detect plagiarism, the suspect document must be compared to document repositories, which we call reference documents, that increases the search space and incurs time penalties. To reduce the search space, we use Latent Dirichlet Allocation (LDA), a probabilistic generative model. LDA consists of a three-level hierarchical Bayesian model and an expectation-maximization (EM) algorithm that exploits Gaussian word distributions (unigram bags of words) for each document to find (n) topics [7].

The pre-processing stage involves converting the source documents into a Bag of Words (PoW) to facilitate the processing process and simplify comparison and calculations. Preprocessing begins by converting the text to lowercase, removing stop words, removing spaces, filtering numbers, unrooting, lemmatizing, removing punctuation, tokenizing each document, and converting plurals to singulars. The process ends with the generation of a bag of unigram words. By a pictorial view Figure 6 illustrates the LDA topic modeling process.

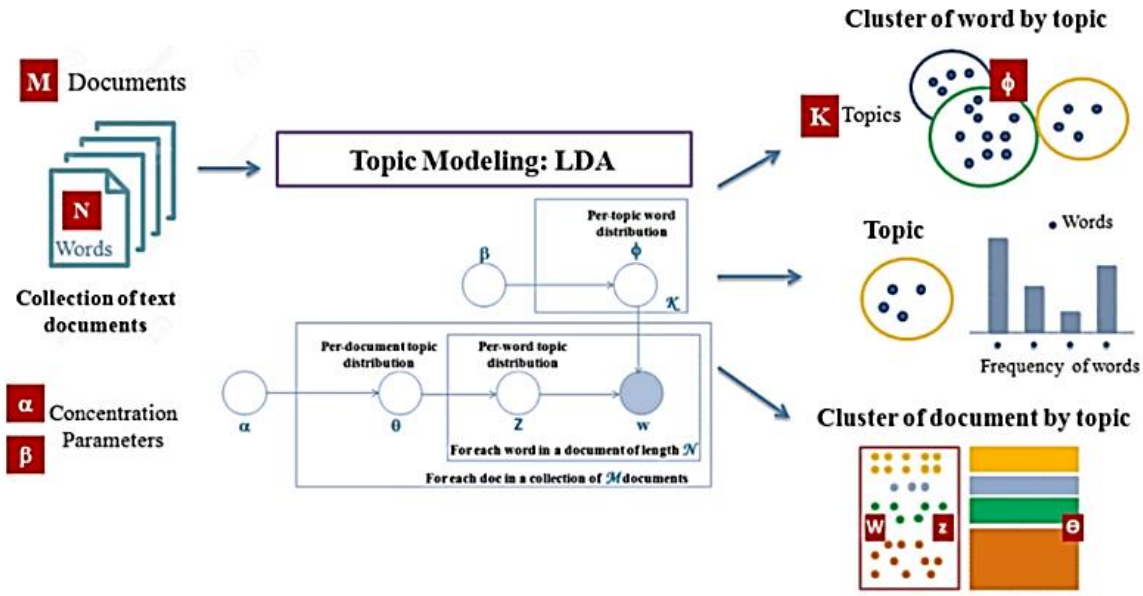

Figure 6: LDA-based Topic Modeling Process LDA.

3.3.3 Stage 3: LDA Documents-Topic Mapping

LDA deliberates a document as a mixture of topics with different probabilities that describe the percentage of the membership of each document for each topic. After obtaining the percentage value for each document, we ranked the documents for the highest value of the topic probability by a matching function. The matching process starts by taking the output of the LDA model and finding the highest percentage for the topic membership value to finally provide a set of classes (topics) as candidate documents for comparison. A simple clarification of the mapping process is introduced in the following counter-example:

Assume document M is "Health insurance companies provide good services to the patients.". The LDA results are = [(0, 0.111817695), (1, 0.5781398), (2, 0.31004253)]. The Mapping Value = Max (Belonging value of n topics) = Max (0.111817695, 0.5781398, 0.31004253) = 0.5781398. Finally, the mapping result for document M is Topic1.

The LDA model is trained using the PAN13 training dataset, and it will be stored for use in the upcoming stage.

3.3.4 Stage 4: Plagiarism Estimation

After generating the bag of words (PoW) for the suspicious document, we need to classify it to a specific topic in accordance with LDA model and the mapping procedure. By specifying the topic of the suspicious document, the candidate documents for later stages are those with the same topic and the search space will be decreased.
The similarity between sentences was measured using Wu & Plamer's algorithm, which gives a score that considers the position of concepts (word and its synonyms) c1 and c2 in the taxonomy relative to the position of the Least Common Subsumer (LCS) (c1, c2). It assumes that the similarity between two concepts in the Wordnet is a function of path length and depth, in path-based measures. The similarity is computed by Equation (1).

The LCS of two nodes *v* and *w*, in a tree or directed acyclic graph (DAG) T, is the lowest (i.e., deepest) node that has both *v* and *w* as descendants, where we define each node to be a descendant of itself (so if v has a direct connection from *w*, *w* is the least common ancestor).

$$\mathrm{Sim}_{\mathrm{wup}}(c1, c2) = \frac{(2 * \mathrm{Dep}(\mathrm{LCS}(c1,c2)))}{(\mathrm{Len}(c1,c2) + 2 * \mathrm{Dep}(\mathrm{LCS}(c1,c2)))} \quad (1)$$

Where, the LCS (c1, c2) = Lowest node in the hierarchy that is a hypernym of c1, c2.

The degree of similarity is a fuzzy measurement issue that depends on many factors, such as the used training set. Alzahrani and Salim [4] suggested a fuzzy indication for plagiarism by setting a threshold value of similarity of 50% as shown in Figure 7. The estimation of plagiarism is based on the threshold value of the similarity measure between a suspect document and some corpus. The specification of this threshold value is related to the nature of the content of the suspect document and differs from one academic organization to another. The final plagiarism model which combines all previous stages is shown in Figure 8.

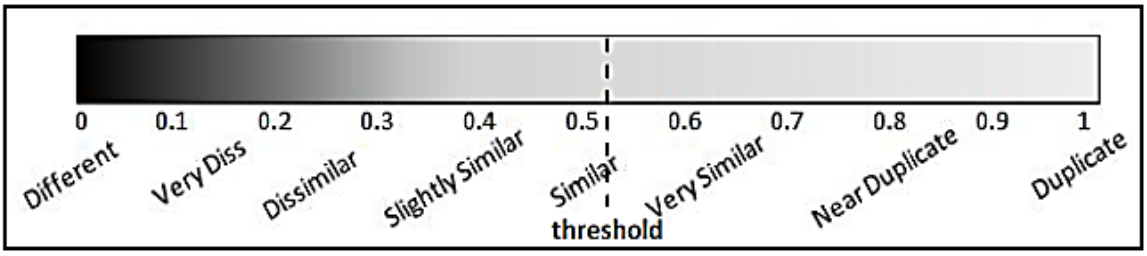

Figure 7: Fuzzy similarity measures with vague boundaries [4]

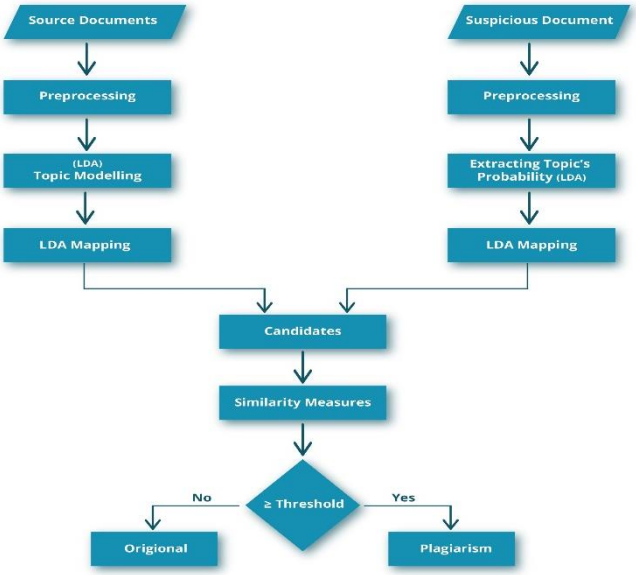

Figure 8: The flowchart of the plagiarism model

4 Results and Evaluation

4.1 Testing

Output data and live examples are used for testing. We trained the LDA with 100 source documents and used 25 suspicious documents for testing. The plagiarism threshold value was set to 75%, so if the similarity measure is higher than this value, we consider the document as plagiarized. Figure 9 presents an example of the result.

```
====================PLAGIARISM DETECTION RESULTS=====================================
||Sus. Document:  Doctors say that brocolli is good for your health.
||Candidate Document:  Brocolli is good to eat. My brother likes to eat good brocolli, but not my men.
||Not Plagiarised
||--------------------------------------------------------------------------------
||Sus. Document:  Doctors say that brocolli is good for your health.
||Candidate Document:  I often feel pressure to perform well at school, but my mother never seems to drive my brother to do better.
||Somewhat Similar
||--------------------------------------------------------------------------------
||Sus. Document:  Doctors say that brocolli is good for your health.
||Candidate Document:  Health professionals say that brocolli is good for your health.
||Plagiarised
||--------------------------------------------------------------------------------
================================================================================
```

Figure 9: Sample result

4.2 Accuracy Measurements

Measurements for the data tested focus on calculating the accuracy, recall, precision, and F-measure.

- Precision is used to show how accurate the model is by determining how many predicted positives are positive. Precision is computed by Equation (2).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

- Recall is used to show how many actual positive classifications were labeled by the model as positive which is evaluated based on Equation (3).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

- F-measure combines the two mesures to provide more powerful and accurate results on model accuracy. F-Measure is computed based on Equation (4).

$$F\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

- Precision is simply a ratio of correctly predicted observations to total observations. Accuracy is an excellent measure, but only when you have symmetrical data sets where the values of false positives and false negatives are nearly the same. Therefore, you need to look at other metrics to evaluate the performance of your model. For our model, we obtained an accuracy of 85%. The accuracy formula is presented in Equation (5).

$$Accuracy = \frac{True\ Postive + True\ Negative}{True\ Postive + True\ Negative + False\ Postive + False\ Negative} \quad (5)$$

The results of these measurements are summarized in Table 2 below. The pictorial view of the measurements in Table 2 is shown in Figure 10.

Table 2: Results summary

Total Precision	0.77
Total Recall	0.83
Total F-measure	0.80
Accuracy	0.85

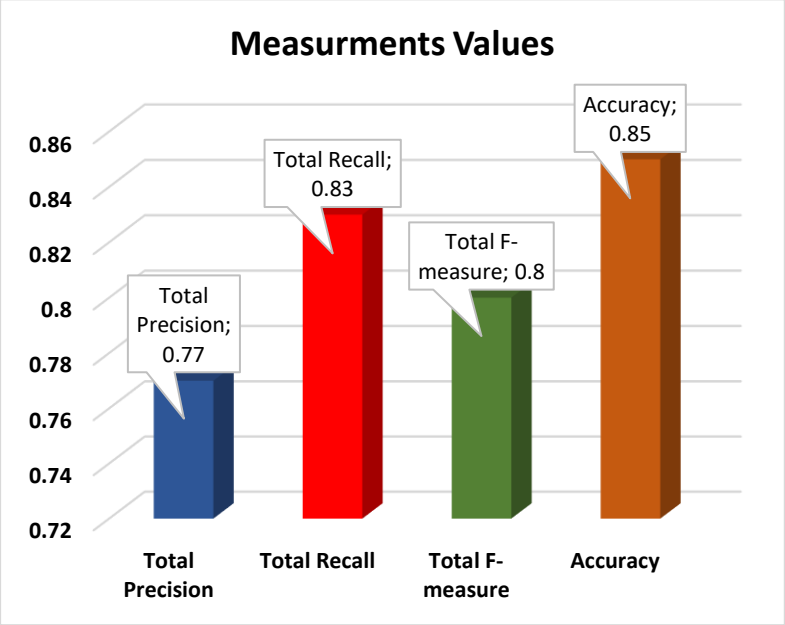

Figure 10: Results summary

5 Conclusion and Future Work

In this research, we developed a plagiarism detection model aimed at discovering extrinsic semantic and syntactic types using LDA to reduce the search space. The approach adopted is based on a large English Wordnet lexical database, LDA and Wu & Plamer's algorithm. The PAN13 dataset was used for training and testing purposes with 100 source documents and 25 suspect documents. The proposed approach was able to find the similarity between the documents. From the results of our experiments, we found that our model can detect syntactic and semantic plagiarism with an accuracy of 85%.

We hope to extend our system to detect more types of plagiarism and use different similarity measures. We also hope to extend the system to detect other languages such as Latin languages (French, Spanish, Italian) and Semitic languages like Arabic.

ACKNOWLEDGEMENTS

The authors extend their appreciation to the Arab Open University for funding this work.

References

[1] Abdi, S. M. Shamsuddin, N. Idris, R. M. Alguliyev, R. M. Aliguliyev (2017). A linguistic treatment for automatic external plagiarism detection, *Knowl.-Based Syst*., 135–146.

[2] K. Aldebei, X. He, W. Jia, J. Yang (2016). Unsupervised multi-author document decomposition based on hidden Markov model, 54th Annual Meeting of the Association for Computational Linguistics, 706–714.

[3] E. S. Al-Shamery, H. Q. Gheni (2016). *Plagiarism detection using semantic analysis*, Indian J. Sci. Technol., 9 , 1–8.

[4] S. Alzahrani, N. Salim (2010). Fuzzy semantic-based string similarity for extrinsic plagiarism detection, Braschler and Harman, 1176 , 1–8.
[5] Bensalem, P. Rosso, S. Chikhi (2019). On the use of character n-grams as the only intrinsic evidence of plagiarism, Lang. Resour. Eval., 53, 363–396.
[6] D. R. Bhalerao, S. S. Sonawane (2015). A survey of plagiarism detection strategies and methodologies in text document, Int. J. Sci. Eng. Technol. Res., 4.
[7] D. M. Blei, A. Y. Ng, M. I. Jordan. (2003). *Latent dirichlet allocation*. J. Mach. Learn. Res., 3, 993–1022.
[8] T. W. S. Chow, M. K. M. Rahman (2009). Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection, IEEE Trans. Neural Netw., 20, 1385–1402.
[9] F. Duarte, D. Caled, G. Xexéo (2017). Minmax circular sector arc for external plagiarism's heuristic retrieval stage, Knowl.-Based Syst., 137, 1–18.
[10] T. A. E. Eisa, N. Salim, S. Alzahrani (2015). Existing plagiarism detection techniques: a systematic mapping of the scholarly literature, Online Inf. Rev., 39, 383–400.
[11] M. Elamine, S. E. Mechti, L. H. Belguith (2017). Intrinsic detection of plagiarism based on writing style grouping, International Workshop on Language Processing and Knowledge Management.
[12] M. Fan, W. Lin, Y. Feng, M. Sun, P. Li (2018). A globalization-semantic matching neural network for paraphrase identification, In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2067–2076.
[13] L. Gang, Z. Quan, L. Guang (2018). Cross-language plagiarism detection based on WordNet, In Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence, 163–168.
[14] Gipp, J. Beel (2010). Citation based plagiarism detection - a new approach to identify plagiarized work language independently, In Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, 273–274.
[15] Gipp (2014). Citation-based plagiarism detection, Springer Fachmedien Wiesbaden, 57–88.
[16] G. Helgesson, S. Eriksson (2015). Plagiarism in research, Med. Health Care Philos., 18, 91–101.
[17] O. Hourrane, E. H. Benlahmar (2017). Survey of plagiarism detection approaches and big data techniques related to plagiarism candidate retrieval, In Proceedings of the 2nd International Conference on Big Data, Cloud and Applications, 1–6.
[18] L. L. Kong, Z. Y. Han, H. L. Qi, M. Y. Yang (2019). Source retrieval model focused on aggregation for plagiarism detection, Inf. Sci., 503, 336–350.
[19] N. Meuschke, B. Gipp, C. Breitinger (2012). CitePlag: a citation-based plagiarism detection system prototype, In Proceedings of the 5th International Plagiarism Conference, 1–10.
[20] R. R. Naik, M. B. Landge, C. N. Mahender (2019). Word level plagiarism detection of marathi text using N-Gram approach, International Conference on Recent Trends in Image Processing and Pattern Recognition, 14–23.
[21] M. Potthast, M. Hagen, T. Gollub, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos, B. Stein (2013). Overview of the 5th international competition on plagiarism detection, In CLEF Conference on Multilingual and Multimodal Information Access Evaluation, 301–331.
[22] M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeno, P. Rosso (2009). Overview of the 1st international competition on plagiarism detection, 3rd Workshop on

Uncovering Plagiarism, Authorship and Social Software Misuse, 1–9.
[23] S. Rakian, F. S. Esfahani, H. Rastegari (2015). A Persian fuzzy plagiarism detection approach, J. Inf. Syst. Telecommun., 3, 182–190.
[24] S. Soleman, A. Fujii (2017). Toward plagiarism detection using citation networks, In 12th International Conference on Digital Information Management, 202–208.
[25] Sornsoontorn, S. Rimcharoen, N. Leelathakul, A. Kawtrakul, P. Ratanaworabhan (2017). Using document classification to improve the performance of a plagiarism checker: a case for Thai language documents, In 21st International Computer Science and Engineering Conference, 219–223.
[26] J. F. A. Traniello, T. C. M. Bakker (2016). Intellectual theft: pitfalls and consequences of plagiarism, Behav. Ecol. Sociobiol., 70, 1789–1791.
[27] K. Vani, D. Gupta (2017). Text plagiarism classification using syntax based linguistic features, Expert Syst. Appl., 88, 448–464.
[28] K. Vani, D. Gupta (2018). Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: comparisons, analysis and challenges, Inf. Process. Manag., 54, 408–432.
[29] V. Vysotska, Y. Burov, V. Lytvyn, A. Demchuk (2018). Defining author's style for plagiarism detection in academic environment, In IEEE 2nd International Conference on Data Stream Mining & Processing, 128–133.
[30] V. Vysotska, O. Kanishcheva, Y. Hlavcheva (2018). Authorship identification of the scientific text in Ukrainian with using the lingvometry methods, In IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, 34–38.
[31] Kaur, M., Gupta, V., & Kaur, R. (2023). Semantic-based integrated plagiarism detection approach for English documents. *IETE Journal of Research*, *69*(9), 6120-6136.
[32] AL-Jibory, F. K. (2021). Hybrid system for plagiarism detection on a scientific paper. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(13), 5707-5719.
[33] Arabi, H., & Akbari, M. (2022). Improving plagiarism detection in text document using hybrid weighted similarity. *Expert Systems with Applications*, *207*, 118034.
[34] Nahar, K. M., Abu Shquier, M., Al-Khatib, W. G., Al-Muhtaseb, H., & Elshafei, M. (2016). Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition. *International Journal of Speech Technology*, *19*, 495-508.
[35] Alsharman, N., Masadeh, R. M., Almomani, O., & Bani-Hani, N. (2023). High-Performance Computing of Building The Dependency Trees and Calculating Tree Edit Distances For Text Similarity. *International Journal of Advances in Soft Computing & Its Applications*, *15*(1).