Turkish Journal of Computer and Mathematics Education Vol.12 No 13 (2021), 5141-5153

Research Article

Hybrid System for Plagiarism Detection on A Scientific Paper

Farah K. AL-Jibory, Mohammed S. H. Al- Tamimi*

a Post graduate Student, Dept. of Computer Science, College of Science, University of Baghdad, IRAQ

Abstract: Plagiarism Detection Systems are critical in identifying instances of plagiarism, particularly in the educational sector whenever it comes to scientific publications and papers. Plagiarism occurs when any material is copied without the author's consent or attribution. To identify such acts, thorough knowledge of plagiarism types and classes is required. It is feasible to detect several sorts of plagiarism using current tools and methodologies. With the advancement of information and communication technologies (ICT) and the availability of online scientific publications, access to these publications has grown more convenient. Additionally, with the availability of several software text editors, plagiarism detection has become a crucial concern. Numerous scholarly articles have previously examined plagiarism detection and the two most often used datasets for plagiarism detection, WordNet and the PAN Dataset. The researchers described verbatim plagiarism detection as a straightforward case of copying and pasting, and then shed light on clever plagiarism, which is more difficult to detect since it may involve original text alteration, borrowing ideas from other studies, and Other scholars have said that plagiarism can obscure the scientific content by substituting terms, deleting or introducing material, rearranging or changing the original publications. The suggested system incorporated natural language processing (NLP) and machine learning (ML) techniques, as well as an external plagiarism detection strategy based on text mining and similarity analysis. The suggested technique employs a mix of Jaccard and cosine similarity. It was examined using the PAN-PC-11 corpus. The proposed system outperforms previous systems on the PAN-PC-11, as demonstrated by the findings. Additionally, the proposed system obtains an accuracy of 0.96, a recall of 0.86, an F-measure of 0.86, and a PlagDet score of 0.86. (0.86). 0.865 and the proposed technique is substantiated by a design application that is used to detect plagiarism in scientific publications and generate non-medication notifications. Portable Document Format (PDF) .

Keywords: Natural language processing, Machine Learning, text mining technic ,External plagiarism detection, Plagiarism detection.

1. Introduction

Plagiarism is a complicated and ethically difficult issue that simply refers to the act of stealing and publishing another author's work under one's own name without recognizing the original author **(Miguel.R 2015)**. Plagiarism is a form of fraud. Authors should properly recognize sources in order to adhere to ethical standards, and plagiarism is a failure to do so. However, the writers' pupils occasionally fail to properly credit the source. These problems are primarily the result of a lack of information on correct citation usage. Thus, plagiarism should be avoided to maintain ethics (**Miguel.R**.**2006)** Perhaps the best definition of plagiarism is "the unacknowledged copying of papers or programs**(Asif.E.etal.,2012)**. Thus, it is necessary to be resolute in one's resistance. Plagiarism, on the other hand, isn't only a problem in academia; it affects nearly every industry. Plagiarism can happen by mistake, but most of the time it is the result of a deliberate procedure **(Durga & Venu 2014)**. The problem of plagiarism has lately been more prevalent as a result of the digital era of materials available on the World Wide Web (WWW). Plagiarism detection (PD) in natural languages using statistical or automated approaches began in the 1990s, with investigations on copy detection mechanisms in digital texts as a forerunner **(Methieu & Michal 2008)**. Since the 1970s, investigations to identify computer code plagiarism in the Pascal and C languages have been conducted to identify code clones and software abuse **(Xie.R 2018)**. To prevent plagiarism, a huge number of researchers have spent decades developing software detection systems **(Hussaim & Dhrub 2018)** . Initially, plagiarism was identified manually (by hand) or through resemblance to previously consulted content. Today, the abundance of available internet materials makes manual detection more difficult. As a result, the development of automatic plagiarism detectors is critical**(Mayank & Dilip 2017) (Efstathio.S .2011)**

.

2.Review Of Related Studies

In(Parth Gupta. et al., 2011). The proposed system focuses on the importance of paraphrases in detecting plagiarism, both monolingually and cross-lingually. To investigate the detection challenges, The authors examined the efficacy of an external plagiarism detection system based on the Vector Space Model (VSM) on the PAN-PC-2011 corpus. The system employed only 250 documents as candidate documents and 20 documents as suspect documents. And the outcome of Monolingual Simulated Plagdet Score (0.0524298), Recall (0.0293390), Precision (0.3780321), and Granularity (1.0541872), and When used in conjunction with any synonym addition mechanism, such as the thesaurus, dictionary, or wordnet, this strategy may be more effective.

In (Asif Ekbal .et al., 2012) offer a method for detecting external plagiarism based on the classic VSM and n-gram language model techniques. The proposed system's methodology is comprised of four major components. In the first step, all texts are processed to create tokens and lemmas, as well as to identify Part-of-Speech (PoS) classes, character offsets, sentence numbers, and Named-Entity (NE) classes. The documents are then forwarded to the pipeline's second stage. Select a subset of documents that may be potential sources of plagiarism in the

second phase. The third phase is a graph-based technique to identify portions that are comparable in the suspect document and selected source documents. Finally, filter out false detections to improve performance; the system is unable to detect cases of translation and has a low recall value; the algorithm, which is based on n-grams, is unable to detect plagiarized cases where common n-grams are insufficient; and the number of documents uses a subset of 1,000 suspicious documents from the PAN. Precision is (65.93), recall is (19.04), granularity is (1.03), and the Plagdet Score is (28.91).**In (Filip Cristian , B. et al., 2013).** proposed a system to detect external plagiarism with the use of Authentic Cop, with the objective of detecting instances of plagiarism in Computer Science academic writings and purposes, the authors experimented with the cosine similarity and term frequency-inverse document frequency (tf-idf)weighting schemes on the PAN 2011 corpus and 1000 randomly selected documents, the authors performed preprocessing, removing stop words, and applying stemming, the authors performed preprocessing, removing stop words, The tested whether there is a distinction between plagiarized and non-plagiarized passages in these circumstances, i.e. if there is a threshold over which the majority of pairs with similarity greater than are plagiarized but the majority of pairs with similarity less than are not, and that are highly similar under the cosine similarity with tf-idf weighting, the prosed system cannot compare in detail the The candidate selection phase will retrieve only the documents from which the current text is most likely to plagiarize; the proposed system tested N-gram sizes of 3, 4, and 5 atoms and discovered that using 4-grams produces the best results in the absence of normalization and stemming; the proposed system demonstrated its effectiveness by performing an application, yielding Precision (0.7609), Recall (0.3377), and Granularity (1.2653). (0.3965). The algorithms used in the candidate selection and detailed analysis phases should be further benchmarked for other combinations of threshold values, and the program should be enhanced to handle other document formats, be accessible via a user-friendly web interface, and include semantic analysis components and stemming support. **In (Asad Abd . et al .,2015)** A method for detecting external plagiarism has been presented. By integrating semantic relationships between words and their syntactic composition, this method can improve the performance of plagiarism detection by avoiding selecting source text sentences that are highly similar to suspicious text sentences but have a different meaning. This system was tested on the PAN-PC-10 and PAN-PC-11 datasets using stop words extracted from the English language.To test and compare the proposed technique's performance, we applied the approach to 200 previously used suspect documents and source documents using four different standard metrics (macro-average Precision, Recall, F-measure, and granularity). The outcome of Plagiarism Detection Using Linguistic Knowledge (PDLK) on PAN-PC-11 systems is as follows: the PDLK has a precision of (0.902), a recall of (0.702), an F-measure of (0.790), and a plagdet of (0.789). However, this result is based on just 200 documents out of 22000 documents, indicating that it does not operate on all datasets.**(Mansi Sahi &Vishal Gupta 2017)** proposed an approach for detecting plagiarism that makes use of both syntactic and semantic information. The three critical stages of the system are preprocessing, deep analysis, and postprocessing. At the preprocessing stage, the documents are segmented, stop words are deleted, stemming is performed, and other operations are performed. The hypothesized document is compared to the original manuscript using various weights for linguistic variables such as local density, inverse path length, depth estimate, and depth feature throughout the entire comparison step. Finally, during the post-processing step, non-plagiarized phrases are filtered away. Additionally, Sahi and Gupta analyzed the precision, recall, accuracy, F-measure, and Plagiarism Detection (PlagDet) score of plagiarism detection algorithms. The system was evaluated using a 200-document PAN-PC-11 standard dataset, which indicates that it will not work with other data sets. On 200 documents, the system obtained Precision (0.949), Recall (0.715), and F-measure (0.815). The system does not perform exhaustive weighing of all potential combinations of linguistic feature functions for the purpose of doing in-depth research on the system.(**Asad Abdi .et al., 2017)** The suggested system introduces an External Plagiarism Detection System (EPDS) that makes use of the Semantic Role Labeling (SRL) approach, as well as semantic and syntactic information. The suggested method is capable of detecting several types of plagiarism, including exact verbatim copying, paraphrase, phrase transformation, and word structure modification. The suggested algorithm operates on the English-language portion of the data set, analyzing 800 questionable papers and their related originals. These papers are divided into two different datasets at random (training and test dataset). The training data set has 450 papers and 350 test documents. Precision (0.921), Recall (0.622), F1 (0.743), Plagdet (0.737), and Granularity (0.737) are the results of the assessed method on the PAN-PC-11 dataset (1.011). The amount employed in testing and training is little, as the total number of English books is 22,000, and the system and system do not assess the effect of stop words on text relevancy.(**Lovepreet Ahuja. et al., 2020)** makes use of semantic information to detect duplicated material in the absence of human intervention; the suggested system makes use of an extrinsic plagiarism detection methodology that is cognitive in nature. The system determines the semantic similarity between two phrases using the Dice measure, which is backed by a lexical database such as WordNet. It also makes use of linguistic characteristics such as path similarities and depth calculation to determine the similarity between two words, which are blended using different weights. It is capable of detecting restructure, paraphrase, exact copying, and synonymized plagiarism, among other things. The PAN-PC-11 corpus was used to assess it. The proposed system is expected to produce results equivalent to or slightly better than those produced by existing systems, and source articles are

preprocessed using NLP features. The system preprocessed the data by normalizing the text, segmenting it into sentences, removing stop-words, and lemmatizing the words, using the English language only. The system's precision (0.934), recall (0.861), and F1-measure (0.875) values are as follows (0.875). A drawback of the proposed approach is its inability to detect complicated examples of textual plagiarism, such as input text summary information and translated text. This method is unable to detect more intricate instances of plagiarism that are manually altered. Because the system is limited to plain text, it cannot identify plagiarism in figures, captions for figures, or flow charts, and it can not use Machine Learning (ML) approaches to identify plagiarism in text documents.

3.Objectives Of The Study

- To make a balance between the two important factors, time and precision in a parallel way when building the algorithm for PD.
- To creating an application that is free for the user to detect plagiarism and is supported with user interfaces and supports the creation reports for the results in a non-modifiable manner.
- To creating adaptive gram in application to solve the problem of N-gram that need higher consumer time.
- To Examination of scientific papers based on the dataset by building a good idea of algorithm that work in PD.

4. Proposed System

The suggested system is a hybrid system with the objective of developing a good algorithm for identifying plagiarism using the PAN-PC-2011 dataset and providing a free application to aid users in detecting plagiarism. The suggested approach is based on data mining techniques and detects plagiarism by comparing suspect and source papers using Jaccard similarity.

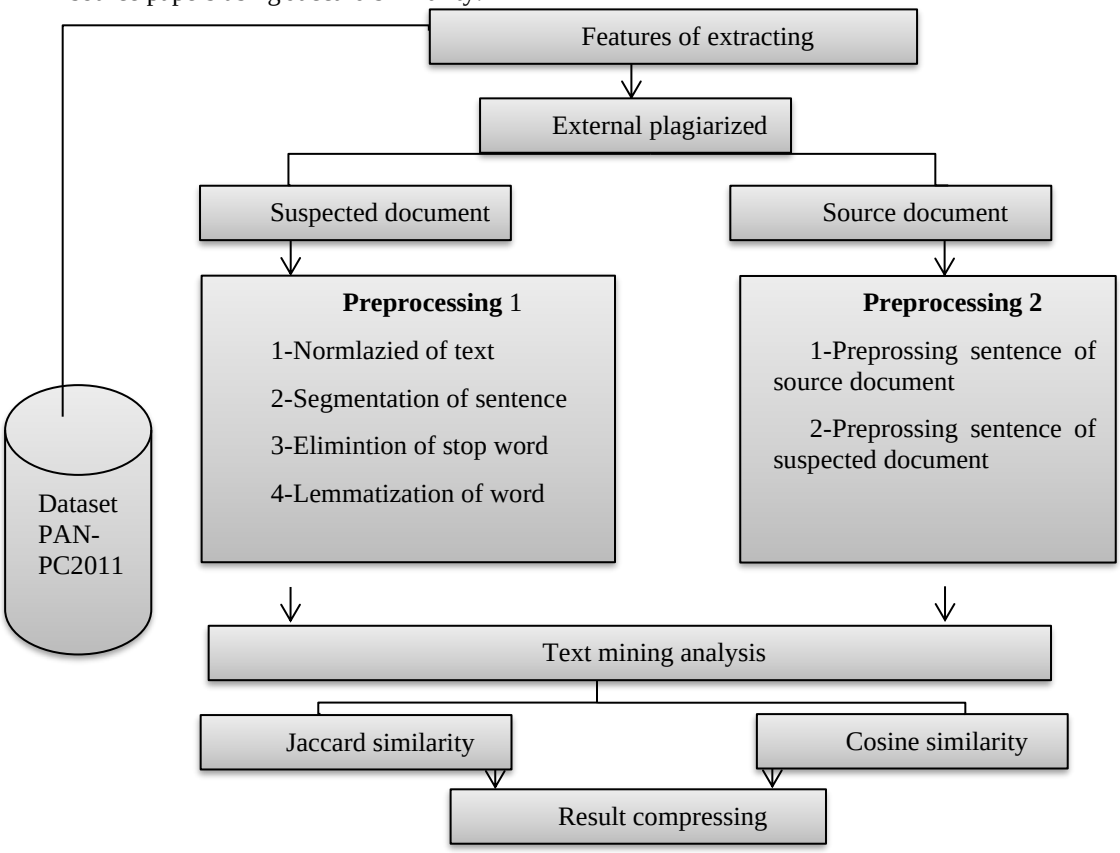


Figure.1 Structure for Proposed Hybrid PD system

5.Dataset Used

The system under consideration currently working on the PAN-PC-2011**(Pan.D 2011) (Salh.A 2011)**. The data set is a corpus that contains documents that have been mechanically and manually plagiarized, and the source documents are texts in the corpus that are based on Project Gutenberg books (www.gutenberg.org). It is based on 22,000 English-language books, 520 German-language novels, and 210 Spanish-language novels. This corpus contains no instances of genuine plagiarism. By contrast, every instance of plagiarism marked in this corpus is

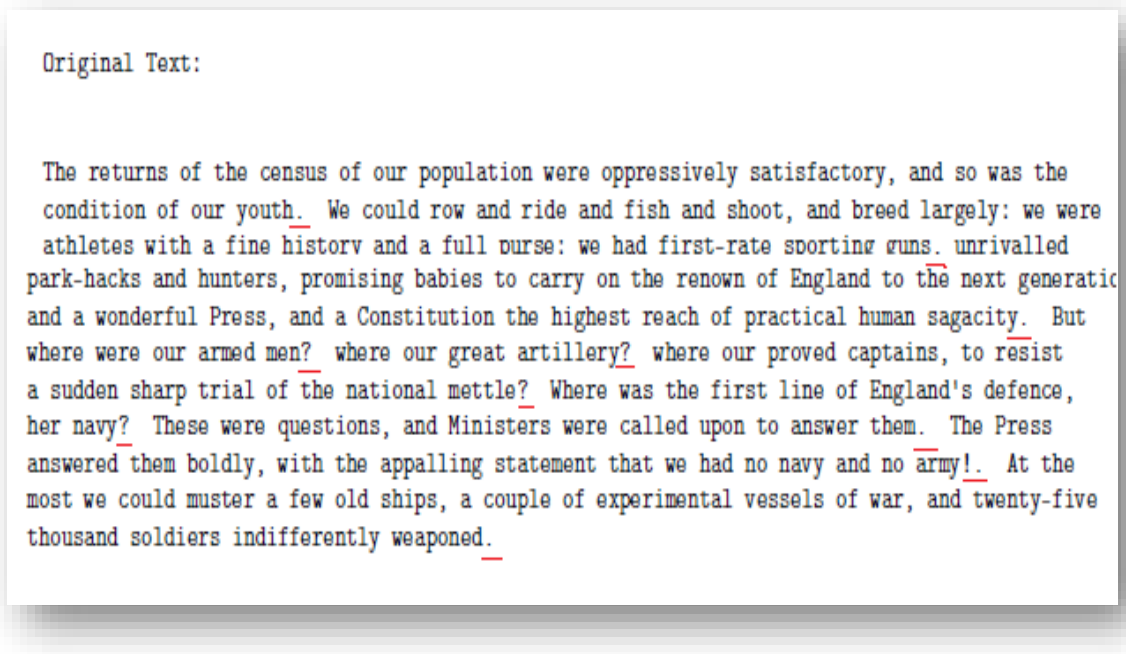
either artificial, i.e., generated by a computer program, or simulated, i.e., purposely created by a human who has committed plagiarism. Additionally, our suggested approach is capable of detecting external plagiarism, which means Given a set of suspicious papers and a set of source documents, the aim is to identify all plagiarized text passages in the suspicious documents and their associated source document text passages. The suggested system is written in English and makes use of approximately 11 folders from a total of 23, with each folder containing 1000 items; hence, the system operates on 11 suspected folders and 11 source folders.

6.Text Mining and NLP

Mining of textual data **(Florin.G 2011)**. (affectionately referred to as data mining) is an Artificial Intelligence (AI) technology that makes use of natural language processing (NLP) **(Salha.A.etal ,2015)** to convert unstructured data in documents and databases into standardized, structured data suitable for analysis or driving machine learning (ML) algorithms. It is the process of sifting through massive collections of papers in order to unearth new information or to aid in the resolution of particular research challenges. It unearths facts, connections, and assertions that would otherwise be lost in the sea of textual big data. Once the data is obtained, it is converted to a structured format that may be further examined or displayed immediately through clustered HTML tables, mind maps, and charts, among other forms. Text mining is a methodology for processing text that utilizes a range of approaches, one of the most prominent being natural language processing (NLP). NLP allows robots to "read" text (or other inputs such as voice) by simulating the capacity of humans to understand a natural language such as English, Spanish, or Chinese. Natural Language Processing (NLP) is a broad term that includes both natural language comprehension and natural language generation (NLG). Natural language processing systems of the contemporary day are capable of evaluating an endless quantity of text-based data in an impartial, consistent way without growing exhausted. They are capable of grasping ideas embedded in complicated settings and interpreting language ambiguities in order to extract critical information and connections or to offer summaries. Automation has become crucial for quickly processing text-based data; the suggested system makes use of NLP and text mining concepts, with the system's mean based on the similarity of text in (source and suspect).

7. Prepressing

The suggested methodology preprocesses suspicious and source documents by enforcing natural language processing features on them. The texts were preprocessed by segmenting them into sentences, normalizing them, deleting stop words, and lemmatizing the terminology for the proposed system. Additionally, there is the following. Figure 2. A portion of the data set's text describing the pre-pressing procedure.


Original Text:

The returns of the census of our population were oppressively satisfactory, and so was the condition of our youth. We could row and ride and fish and shoot, and breed largely: we were athletes with a fine history and a full purse: we had first-rate sporting guns, unrivalled park-hacks and hunters, promising babies to carry on the renown of England to the next generatic and a wonderful Press, and a Constitution the highest reach of practical human sagacity. But where were our armed men? where our great artillery? where our proved captains, to resist a sudden sharp trial of the national mettle? Where was the first line of England's defence, her navy? These were questions, and Ministers were called upon to answer them. The Press answered them boldly, with the appalling statement that we had no navy and no army!. At the most we could muster a few old ships, a couple of experimental vessels of war, and twenty-five thousand soldiers indifferently weaponed.

Figure 2. Sample of original text in data set

7.1 Segmentation

The content of the suspect and source documents has been broken down in phrases. Both pieces of information are then expressed in a series of sentences. Text fragments are denoted by the ". ", "?", and "!" symbols. The

segmentation of the original text is seen in Figure 3.

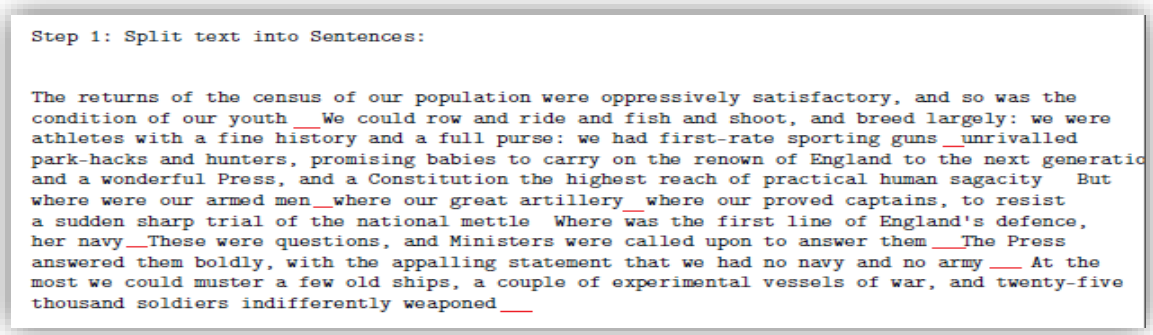


Figure 3.segmentation step

7.2 Normalization

When comparing documents, several minor characters in the text may be ignored. The system's efficiency can be boosted by excluding specific letters from the text. Commas, colons, semicolons, brackets, special characters, quotations, and white spaces are not required for similarity evaluation. Furthermore, the term may be truncated or have an alternate spelling that must be standardized. Following segmentation, Figure 4 normalizes the text.

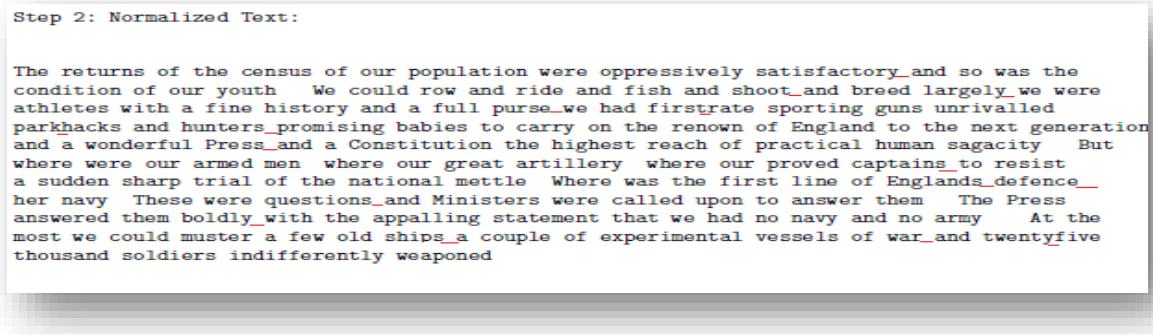


Figure4.norlmalized the text after segmentation.

7.3 Eliminating stop-words

Articles, propositions, and conjunctions are commonly used throughout the text. Stop words account for around 40% to 50% of all words in ordinary text **(Lisn.Z 2016)**. Furthermore, these are useless terms. By excluding certain phrases from the text, the system's computation time is reduced while its efficiency and quality are boosted. The suggested technique removes stop-words from the list of stop-words in the Natural Language Toolkit (NLTK). The list contains around 160 stop-words. (as seen in Figure 5) After text normalization, stop-words are eliminated.

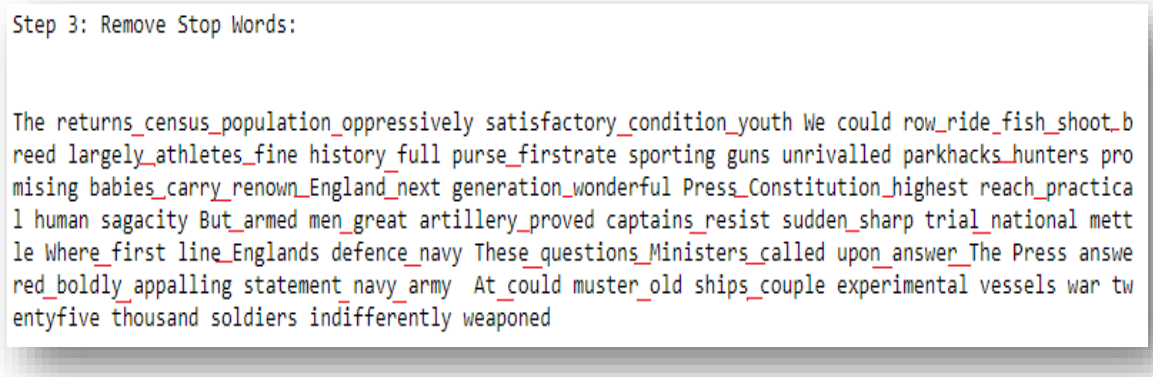


Figure5. Eliminating stop-words

7.4 Lemmatization

Is a word-processing approach that utilizes lexical and morphological interpretation to eliminate superfluous ends and generate the dictionary base form of a word, referred to as a lemma. By reverting to the words' dictionary definitions, meaningful comparisons become possible. The suggested system makes use of lemmatize to execute word lemmatization, which is accomplished through the use of a wordnet, and Figure 6 represents the final stage of preprocessing in which the lemmatized text is shown.

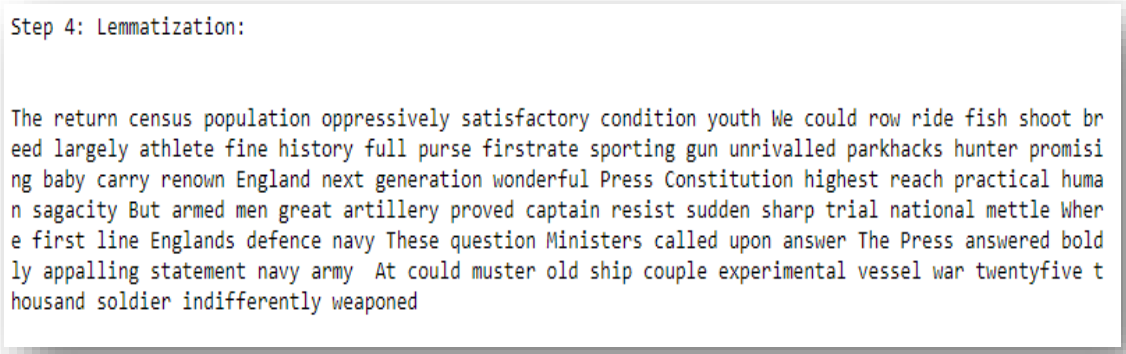


Figure6.lemmatization step

8.Jaccard and Cosine Similarity Measures

which are advantageous for sparse data such as texts, and the suggested approach made use of jaccard and cosine similarity **(Benedikt.w.etal ,2016)(Pang.N.etal ,2011)**. In this example, similarity is generally defined as 1 when attribute values match and 0 when they do not match. A dissimilarity is defined in the opposite manner: 0 indicates that the attribute values match, while 1 indicates that they do not.

8.1 Jaccard Measure

Assume that x(source document) and y(suspect document) are two data objects representing two rows (two transactions) of a transaction matrix. If each asymmetric binary attribute is associated with a document in a dataset, a value of 1 indicates that the document was purchased, while a value of 0 indicates that it was not purchased. The following equation represents the Jaccard coefficient equation **(Shiling.S.etal,2011)**, which is frequently denoted as J.

$$J(X,Y)=\frac{|x\cap y|}{|y\cup x|}=\frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|} \quad \ldots\ldots(1)$$

this measure supports the time factor as the time executed faster than cosine measure , the fowling algorithm show the jaccard similarity in proposed system:

Algorithm1. Jaccard Algorithm

Input : Reference R, Suspect S
Output: compute Jaccard similarity (JaccSim)
Begin : **Step 1:** Count only common words in both R and S, which is the intersection between them as follows: D=R∩S **Step 2:** Count all the words appear in both R and S, which is the union of R,S C=R U S – D **Step 3:** The Jaccard similarity index is Output JaccSim=D/C **End**

8.2 Cosine Measure

The cosine similarity is a widely used measure of document similarity. It is a measure of similarity between two vectors calculated by multiplying the cosine angle values of the two vectors being compared. The equation **(florin.G 2011) (Suplakit.N.etal ,2013)**holds true if x and y are two document vectors.

Cosine(x,y)= $x.y \div ||x||*||y|| \quad \ldots\ldots(2)$

where · indicates the vector dot product, $\mathbf{x} \cdot \mathbf{y} = \int_k^n = 1$ Xk Yk, and ||X|| is the length of vector **X**, $||X||=\sqrt{\sum_k^n = 1\ X_k^2} = \sqrt{X.X} \quad \ldots\ldots(3)$

then can be written as Equation $\frac{x}{||x||} * \frac{y}{||y||}$ = X′* Y′……(4)

As a means of increasing efficiency and performance, cosine similarity is used to calculate the number of vector cosine angles in documents and to construct a hierarchical clustering method for them. The fowling method demonstrates cosine similarity in the proposed system:

Algorithm2.Cosine Algorithm

Input : Reference R, Suspect S

Output: compute cosine similarity

Begin :

Step 1: Estimate frequency $f(w_i)$ of each word w_i in Document S

$f_i = \frac{\sum w_j}{N}$

Where N is the total words in source document S

Step 2: Estimate frequency $f(w_j)$ of each word w_j in Document R

$f_j = \frac{\sum w_j}{M}$

Where M is the total words in reference document R.

Step 3:

For i =1 to N

For j=1 to M

if (j == j)

Calculate the distance between similar words using the following formula:

$$\boldsymbol{dist_k = f_i \times f_j}$$

else

$$dist_k = 0$$

k=k+1

End for

End for

Step 4: Cosine similarity is:

$$cosSim = 1 - \frac{\sum_{k=1}^{K} dist_k}{K}$$

End

9. Result Threshold

The threshold to proses in hybrid technic, Assume the threshold of cosine similarity measure is (Tc), and Jaccard index is (Tj), where cosine metric and Jaccard index for sentence (s) are denoted as (Cs) and (Js) respectively and After multiple tests in which the threshold was that used to compare between cosine and Jaccard measurements result in document , the suggested system invents a hybrid structural for cosine and Jaccard measurement based on threshold that based on execution of mathematical model Jaccard = 0.1 to 0.2 and Cosine = 0.3 to 0.6
and the following algorithm is how threshold work in proposed system.

Algorithm 3.Threshold

Input : Cosine similarity , Jaccard similarity

Output: compute adaptive Threshold TJ,TC

Begin

Step 1:

// compute jacquard threshold Tj

X=$\sum_i^M \sum_j^N Jaccard\ sim\ (S,R)$

Y= M*N

Tj= X/Y

//compute mean average for Tj

M=$\sum Tj\ /\ \sum N$

Step 2:

// compute cosine threshold Tc

X=$\sum_i^M \sum_j^N cosine\ \ sim\ (S,R)$

Y= M*N

Tc= X/Y

//compute mean average for Tc

M=$\sum Tc \,/\, \sum N$

End

10. Proposed Hybrid System

A hybrid system uses the following relationship that based on the equation of Jaccard and cosine Measurement and the threshold algorithm that explains in above.

$$Plag = \{s{:}s \in d, \alpha_s \oplus \beta_s = 1\} \ldots..(7)$$

Where (*s)* is the suspicious sentence in document *D*, $\oplus$ is a logical OR operation,

and

$$\alpha_s = \begin{cases} 0, & if\ C_s > Tc \\ 1, & if C_s < Tc \end{cases} \ldots\ldots(8)$$

and

$$\beta_s = \begin{cases} 0, & if\ j_s > Tj \\ 1, & if j_s < Tj \end{cases} \ldots\ldots\ldots(9)$$

In words, if any sentence(*s)* has one similarity measure, or both, exceeds the threshold of each measure is considered a plagiarized sentence. For document *D*. the adaptive threshold work in the following algorithm:

Algorithem4. hybrid system

Input : R,S
//preprocess Refers document , preprocess Suspect document

Output: similarity of document S

Begin
Step1:
For i=1 to N do //compute jacquard similarity
Based on algorithm 1
Jacc sim (R,S)= (R∩S)/(RUS)
End
Step 2 :
For i=1 to N do //compute cosine similarity
Based on algorithm 2
cos sim (R,S)= (R.S)/(||R||*||S||)
End
Step 3:
Based on algorithm 3
if CosSim >TC $\alpha_s = 1$ else $\alpha_s = 0$
if JaccSim >TJ $\beta_s = 1$ else $\beta_s = 0$
step 4:

$$hybridIndex = \alpha_S \oplus \beta_s$$

where $\oplus$ is a logical OR operation
If hybridIndex =1 the document is plagiarism
Else
hybridIndex =0 the document is not plagiarism
End

11.Experiments and Discussion

To facilitate comprehension of the proposed system, researchers will outline its components and compare it to another system's dataset, PAN-PC-2011. The method of plagiarism between the source and suspect documents in the data set will be explained in the same step-by-step fashion as the algorithm above. an all the stages will be run on each item in each folder of the suspect folder that contains the resource folder, so the procedure of each step will be conducted in a one-to-many fashion. The implementation method is to take a document from the suspect folder and pass it on to all the source folders, and this process was used to evaluate the proposed system's algorithm's efficiency and to extract the error rate for the purpose of evaluation in order to obtain an accurate result. The best result of the two measurements used in the proposed system (cosine and Jaccard) is r. When the experiment is conducted on a data set, the hybrid system established on the Jaccard threshold (0.2) and the Cosine

threshold produces the best results (0.5). These are experimentally determined values. The system being proposed Compares the suspect document to all other documents in the reference folders Due to the similarities in the document's names (suspect and resource). If the document's names are similar, they discovered the actual value, which indicates that the document is genuine Plagiarism. The outcome is (actual value), which is then put in a table for use in confusion metrics (zero, one) If the result is (one is true (T)), then the document is plagiarized; if the result is (zero is false (F), then the document is not plagiarized, and the result is stored in the table for use in metrics. After determining the document's real value, we will determine the document's predicate value, which will depend on whether the document is plagiarized (T) or not (F), as determined by the suggested hybrid system, and the resulting predicted value will be the logical value (zero, one) If the result is (one is true(T)), then the document is plagiarized; if the result is (zero is false(F), then the document is not plagiarized, and the result will be stored in the database for use in metrics. And once the actual and predicate values are determined, they are compared between two tables to determine if there is a difference between the two tables recorded in the table. This is an error percentage that is utilized in confusion metrics to determine the measurement that determines the proposed system's efficiency.

12. Evaluation Metrics

The recommended approach for evaluation will be utilized. A matrix of befuddlement**(Pang.N.etal 2011)**. Which will be conducted in error rate and summarizes the number of occurrences properly or wrongly predicted by a classification model. Counts calculated in a confusion matrix are frequently referred to using the following terminology:

a precision: It is quantified by the standard deviation of a collection of data, whereas bias is quantified by the difference between the mean of the collection of data and the known value of the item being quantified. Precision is the percent of records that are genuinely positive in the group that the classifier has declared as a positive class, as defined by the following precision equation **(suplakit.N. etal**, **2013)**

Precision(p) = $\frac{\text{TP}}{\text{TP + FP}}$ …..(9)

By applied the equation (9) for all document we calculate by using mean average the equation is : average precision=$\frac{\sum p}{\sum N} = \frac{52788}{5500} = 0.958$

Recall: Recall quantifies the proportion of positive examples properly predicted by the classification, and its value is identical to the genuine positive rate.

Recall(*r)* = $\frac{TP}{TP + FN}$ ……(10)

By applied the equation (10) for all document we calculate by using mean average the equation is : average precision=$\frac{\sum r}{\sum N} = \frac{5278}{5500} = 0.959$

F1:The harmonic mean of accuracy and recall, as well as the flowing equation, denote F1, and the equation is [29].

F1= $\frac{2rp}{r+p} = \frac{2 \times TP}{2 \times TP + FP + FN}$ …..(11)

By applied the equation (11) for all document we calculate by using mean average the equation is : average precision=$\frac{\sum r}{\sum N} = \frac{4770.7801}{5502} = 0.867$

Plagdet:It is defined as the product of precision, recall, and granularity and is described as follows [30].

plagdet(S, R)=$\frac{\text{F1}}{\log2(1 + \text{Gran(S,R)})}$ …..(12)

The plagdet metric was used to rank the contestants in the PAN plagiarism detection competitions.

By applied the equation (12) for all document we calculate by using mean average the equation is : average precision=$\frac{\sum plagdet}{\sum N} = \frac{4770.7801}{5500} = 0.867$

In this part, we will compare the efficacy of several systems on the PAN-PC-2011 (Sys-P**(Mansi&Vishal 2017)**,Sys-Q**(Asad.A.etal,2017)**,Sys-R**(Asad.A.etal**,**2015),**Sys-S**(Asif.E.etal,2012)**,Synth-Sema **(Lvepreet.A.etal ,2020)**The all systems based on the Evaluation Metrics that explain in above . Table 1 shows the performance evaluation of the the systems based on (Precision, Recall, F1-Score, palgDet)

Table .1. Performance evaluation of different systems in core paper with proposed system.

Name of sys	Precesion	Recall	F1-Score	palgDet
Proposed system	0.959	0.959	0.867	0.867
Synth-Sema (core paper)	0.934	0.861	0.875	0.875

Sys-P	0.956	0.743	0.836	0.836
Sys-Q	0.921	0.622	0.743	0.737
Sys-R	0.902	0.702	0.79	0.789
Sys-S	0.659	0.19	0.295	0.289

And the Figure 8 is Plagiarism Algorithms Comparison

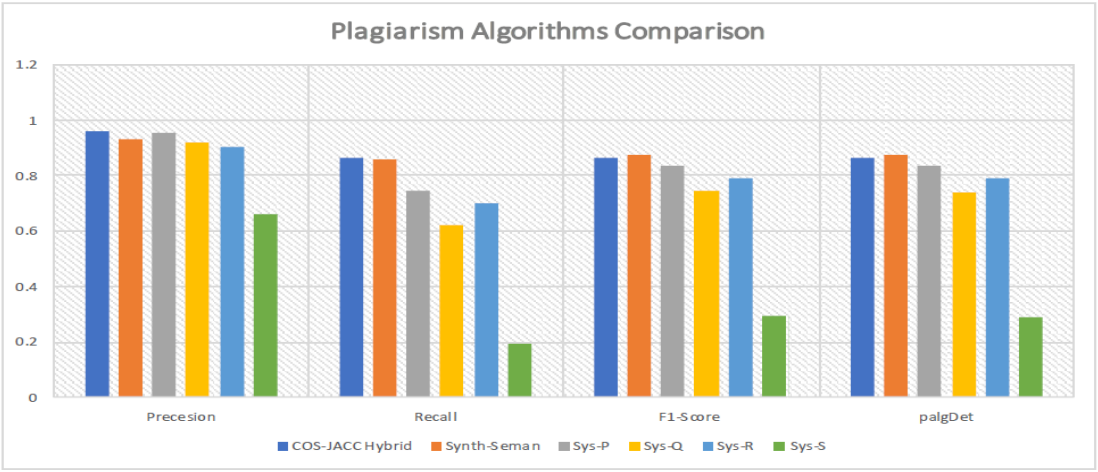


Figure 8. Plagiarism Algorithms Comparison

13. Comparison with Previous Studies

Several PD methods and their findings have been reported in the literature in recent years. The findings of the suggested approach are compared to some of the approaches mentioned in the literature in this section. Table 3.3 compares the detection measurement in the confusion matrix obtained by the proposed method to that obtained in prior investigations. Our proposed strategy appears to be superior to others based on the stated results.

Table2. Comparison with Previous Studies

No.of refrains	dataset	No.of document	percentage
Proposed system	PAN-PC-2011	used 11 folder suspected and 11 folder source Each folder contend 1000 item(doc)	P:0.959~0.96 R:0.959 F1:0.867 Plag:0.867
9	PAN-PC-2011	250 doc source And 20 suspected	P:0.3780321 R:0.0293390

		doc	Plag:0.0524298 G:1.0541872
10	PAN-PC-2011	1000 doc just used suspected doc	P: 0.659 R: 0.196 F1: 0.295 G: not used Plag:0.289
11	PAN-PC-2011	1000 randomly doc	P: 0.7609 R: 0.3377 G: 1.2653 Plag:0.3965
12	PAN-PC-2011 PAN-PC-2010	200 doc in PAN-PC-2011	P: 0.902 R: 0.702 F1:0.790 G: not used Plag:0.789
13	PAN-PC-2011	200 doc in PAN-PC-2011	P:0.949 R:0.715 F1:0.815 G: not used Plag: not used
14	PAN-PC-2011	800 doc of suspect and source document	P:0.921 R:0.622 F1:0.737 G:1.011 Plag:0.737
15	PAN-PC-2011	A few doc and not	P: 0.934

		tell the number	R:0.861 F1:0.875 G:not used Plag:0.875

14. Conclusion and Future Work

The effectiveness of a plagiarism detection system is mostly determined on how the text-processing process occurs within the documents, as well as how to collect raw data free of manipulative manipulation and symbols in order to compare them and reveal the degree of similarity between the source and suspect documents. To ensure a correct comparison process, and Cosine Measurement supports accuracy but not time, and Jaccard Measurement supports time but not accuracy, the hybrid system relied on the threshold that was obtained and estimated based on experience and implemented, and the system used a mathematical model to balance the two. In the future, a method for detecting plagiarism may be developed that utilizes dialect interpreters to identify plagiarism in transformed papers including many languages. A method for detecting plagiarism in figures, flowcharts, translated text, photos, and figure captions may be developed.

Reference

1. Miguel R. D. Ph ,(2015). "Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing,". Office of Research Integrity (ORI), pp. 1–71.
2. Miguel R. D. Ph,(2006). "Avoiding plagiarism , self-plagiarism , and other questionable writing practices : A guide to ethical writing,". Office of Research Integrity (ORI), pp. 1–63.
3. Asif E, Sriparna S, Gaurav C. (2012). "Plagiarism detection in text using Vector Space Model". Proceedings of the 2012 12th International Conference on Hybrid Intelligent Systems, HIS 2012. 978-1-4673-5116-4
4. Durrga B,Venu G,(2014)."UNDERSTANDING PLAGIARISM FOR CONTEXTUAL FEATURES Abstr". International Journal of Software &Hardware Researvhe in Engineering pp. 24–27, 2014.
5. Mathieu F, Michael R ,(2008)."A comparison of common programming languages used in bioinformatics" .BMC Bioinformatics, vol. 9, pp. 1–9.
6. xie R,(2018). "an overview of plagiarism recognition techniques". international journal of knowledge and and language processingc , volume 9, number 2, 2018, 2191-2734
7. Hussain C, Dhruba B,(2018) ." plagiarism: taxonomy, tools and detection techniques". arXiv , ISBN: 978-93-82735-08-3.
8. Parth G, Khushboo S, Prasenjit M, Paolo R, (2011)."Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism". IR-Lab,DA-IICT,India .
9. Mayank A , Dilip S (2016). "A state of art on source code plagiarism detection. ". 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), 978-1-5090-3257-0.
10. Fili B, Adrian S, Traian R, and Razvan R, (2013)."Automatic plagiarism detection system for specialized corpora" . Proc. - 19th Int. Conf. Control Syst. Comput. Sci. CSCS 2013, no. June, pp. 77–82,
11. Mansi S, Vishal G,(2017) ."A Novel Technique for Detecting Plagiarism in Documents Exploiting Information Sources" . Cognitive Computation., vol. 9, no. 6, pp. 852–867, 2017.
12. Asad A, Norisma I, Rasim A, Ramiz A,(2015) ."PDLK: Plagiarism detection using linguistic knowledge". Expert Systems with Applications. Appl., vol. 42, no. 22, pp. 8936–8946, 2015
13. Asad A, Siti S, Norisma I, RasiM A ,(2017) ."A linguistic treatment for automatic external plagiarism detection". Knowledge-Based Syst., vol. 135, no. November, pp. 135–146.
14. Lovepreet A, Vishal G, Rohit K, (2020)."A New Hybrid Technique for Detection of Plagiarism from Text Documents" .Arab. J. Sci. Eng., vol. 45, no. 12, pp. 9939–9952, 2020.
15. https://pan.webis.de/data.html.
16. Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013, March). "Using of Jaccard coefficient for keywords similarity". In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, No. 6, pp. 380-384).
17. Lisna Z,(2016)."Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method" . *Comput. Eng. Appl. J.*, vol. 5, no. 1, pp. 11–18.
18. Elavarasi, S. A., Akilandeswari, J., & Menaga, K. (2014). "A survey on semantic similarity measure".International Journal of Research in Advent Technology, 2(3), 389-398.
19. Pang N (2011). Introduction to data mining.doi:10.1007/978.3-642-197721-5-1.

20. Shiliang S, Chen, J, Junyu C, (2017). "A review of natural language processing techniques for opinion mining systems". Information Fusion, 36, 10–25
21. Gorunescu, F. (2011). Introduction to Data Mining. Data Mining, 1–43
22. SALHA. A,(2012). "Structural Information and Fuzzy Semantic Similarity".Universiti Teknologi Malaysia.
23. Salha A , Naomie S, Vasile P, (2015). "Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model". *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 27, no. 3, pp. 248–268.
24. Trevor C, Chris B, Mirella L(2008) ."Constructing corpora for the development and evaluation of paraphrase systems". *Comput. Linguist.*, vol. 34, no. 4, pp. 597–614.
25. Stamatatos, E. (2011). " Plagiarism detection using stopword n-grams". Journal of the American Society for Information Science and Technology, 62(12), 2512–2527.