International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Detection of Plagiarism in Document Using Cosine Similarity

S. H. Rajput
(Asst. Prof.)
Department of Computer Engineering
SSBT's College of Engineering & Technology. Jalgaon, India
bs.rajput26@gmail.com

Chetan Subhash Nhalade
(BE Computer)
nhaldechetan421@gmail.com

Satyam Sunil Pandhare
(BE Computer)
satyampandhare2331@gmail.com

Pooja Dhanraj Mali
(BE Computer)
poojamaliofficial@gmail.com

Vaishnavi Vilas Gadhe
(BE Computer)
vaishnavigadhe9@gmail.com

Abstract— **The issue of plagiarism has been a significant concern in the academic and professional world for many years. With the increasing availability of digital content, it has become easier for people to copy and paste information from various sources without proper attribution. To address this problem, many methods have been developed to detect plagiarism in documents. One effective method for detecting plagiarism is to use cosine similarity. This technique compares the similarity between two documents based on the angle between their vectors. If the angle is small, the documents are similar, and if the angle is large, they are dissimilar. By applying this method, plagiarism can be detected by comparing the similarities between the suspiciousdocument and a set of reference documents. This paper reviews the use of cosine similarity in detecting plagiarism and its effectiveness. The results show that cosine similarity is a reliable method for detecting plagiarism and can be used in various applications.**

Keywords—plagiarism, detection, cosine similarity.

I. INTRODUCTION

Instantaneous information access made possible by the Internet has also resulted in the emergence of massive amounts of unstructured data, particularly text. The widespread popularity of the practise of copying and reusing content without permission is a significant disadvantage of the simple information access made accessible by the Internet. The unauthorised reuse of ideas or content without giving proper credit to the original authors is known as plagiarism. The threat of plagiarism to academic integrity and authenticity must be appropriately addressed. Plagiarismis a significant issue in academic and professional settings, where individuals may present someone else's work as their own. With the widespread availability of online resources, ithas become easier to access and copy content from various sources. Detecting plagiarism in documents has, therefore, become more crucial than ever before. One of the most commonly used techniques for plagiarism detection is cosine similarity, which measures the similarity between two documents by comparing their word frequency vectors. In this approach, documents are represented as a vector of word frequencies, and cosine similarity is calculated as the cosine of the angle between the two vectors. The closer the angle is to zero, the higher the similarity between the two documents. This method has been shown to be effective in detecting plagiarism, particularly when dealing with large volumes of text. and pasting someone else's work without attribution, to paraphrasing or summarizing someone else's ideas without giving them credit. Plagiarism is a form of cheating, and it can have serious consequences for both the person who commits it and the person whose work is plagiarized.

II. LITERATURE REVIEW

A. *Plagiarism*

Plagiarism originated from Latin "plagiarus" which means kidnapping. The definition of plagiarism according to Big Indonesian Dictionary is "plagiarism that infringes copyright". Meanwhile, according to [1] plagiarism is the act of copying or stealing the others works such as ideas, writing ideas, then claim it as a result of his own work without including reference of the original source. There are several categories of plagiarism, namely word by word plagiarism, Word switch plagiarism, Metaphor plagiarism, Idea plagiarism, and self-plagiarismplagiarism based on the percentage of words taken or traced is divided into 3 categories, such as:

a) Light Plagiarism: < 30%.

b) Medium Plagiarism: 30% - 70%.

c) Heavy Plagiarism: >70%

B. *Text Preprocessing*

Cleaning and preparing text data for subsequent analysis is known as text preprocessing. It involves several steps that help to transform raw text data into a structured format that can be easily analyzed using various machine learning or natural language processing techniques.

The following are some common steps involved in text preprocessing :

1. Tokenization: This involves splitting the text data into individual words or tokens.

2. Lowercasing: Converting all the text data to lowercase helps to eliminate the differences between the same words written in different cases.

3. Stop word removal: Stop words are common words that do not add any significant meaning to the text. These words are removed to reduce the dimensionality of the data and improve the processing time.

4. Stemming and Lemmatization: Stemming involves reducing words to their base form by removing suffixes and prefixes, while lemmatizationinvolves reducing words to their base form based on their morphological analysis.

5. Removing punctuation and special characters: This involves removing all the special characters and punctuation marks from the text data

6. Indexing is a process done to build an index database of document collections.

C. TF-IDF Weighting

TF-IDF (Term Frequency-Inverse Document Frequency) is a weighting scheme used in information retrieval and text mining to evaluate the importance of a term (word or phrase) in a document or corpus of documents. It is a numerical measure that reflects how important a word is to adocument in a collection or corpus.

TF-IDF consists of two main components:

1. Term Frequency (TF): It measures the frequency of a term in a document. It is computed by dividing the total number of terms in the document by the frequency with which each phrase appears.

TF = (Number of times term appears in a document) / (Total number of terms in the document)

2. Inverse Document Frequency (IDF): It measuresthe rarity of a term in a collection or corpus of documents. It is calculated as the logarithm of the total number of documents in the collection divided by the number of documents that contain the term.

IDF = log_e(Total number of documents in the collection / Number of documents that contain the term)

$$Wt,d = TFt,d \times ln\,(\,N\,/df t)$$

where Wt,d is the value of the weight of the word t in document d. The value of TFt,d is the frequency of the word t in document d. N is the total document and dft is a lot of documents containing the word t.

Once the TF and IDF values are calculated, they are multiplied to get the final TF-IDF weight for a term in a document. The higher the TF-IDF weight of a term, the more important it is to the document

D. Cosine Similarity

Cosine Similarity is a method for measuring the level of similarity between two vectors. Calculations in this method are done by calculating the Cosine value between two vectors

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Eaquation I

The cosine similarity between two documents (or text vectors) can be calculated using the following formula:
cosine similarity = (Σi Xi * Yi) / (√(Σi Xi^2) * √(Σi Yi^2))
where Xi and Yi are the TF-IDF weights of the ith term in the two documents, respectively. TF-IDF is a commonly used weighting scheme that measures the importance of a term in a document, based on its frequency in the document and rarity in the corpus.

The formula calculates the cosine of the angle between the two vectors, which reflects the similarity between the two documents. A cosine similarity of 1 indicates that the two documents are identical (i.e., have the same set of terms with the same weights), whereas a cosine similarity of 0 indicates that the two documents are completely dissimilar (i.e., have no common terms).

This formula can be extended to calculate the cosine similarity between multiple documents or text vectors. In this case, the formula can be written as:
cosine similarity = (Σi Xi * Yj) / (√(Σi Xi^2) * √(Σj Yj^2))
where Xi and Yj are the TF-IDF weights of the ith term in the first document and the jth term in the second document, respectively. This formula computes the cosine similarity between all pairs of documents or text vectors in a collection or corpus, and can be used for various clustering or classification tasks.

Overall, the cosine similarity formula is a powerful tool for measuring the similarity between documents or text vectors, and can be used in a variety of research applications that involve natural language processing or information retrieval.

III. RESEARCH METHODOLOGY

A. Data Collection

In this study, the data used is the dataset of research papers. Data obtained through kaggle that located at https://www.kaggle.com/datasets/temuujinerdene/nlpresearchers.

B. Proposed System

The system in the research will be implemented by using Python programming language. The database used isMySQL.

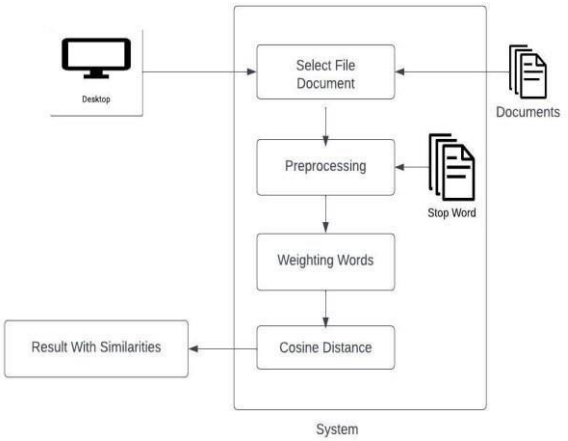
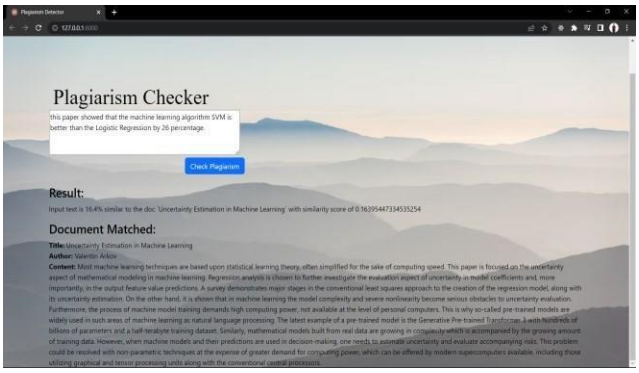


Figure 3.1. Proposed System

IV. RESULT AND DISCUSSION

A. Result

Result after input the text in query format.



Comparative Analysis of Cosine Similarity Calculations – Analysis of cosine similarity calculations is done by comparing manual calculation simulations and calculation simulations in the website. Manual calculations are used as a comparison because manual calculations are calculations that yield true values. This analysis was conducted to determine the level of accuracy of the automatic cosine similarity calculation on website.

TABLE 1 Result of Determining Term Frequency

Words	Input_doc (A)	Output_doc (B)
this	1	3
Paper	1	1
Showed	1	1
The	2	16
Machine	1	6
Learning	1	4
Algorithm	1	0
Svm	1	0
Is	1	6
Better	1	0
Than	1	0
Logistic	1	0
Regression	1	2
By	1	2
26	1	0
Percentage	1	0
That	1	1

Determine the number of intersecting word values, the results at this stage are shown in Table 2.

TABLE 2 Result of Determining the Number of Intersecting Word Values

Words	A X B
this	3
Paper	1
Showed	0
The	32
Machine	6
Learning	4
Algorithm	0
Svm	0
Is	6
Better	0
Than	0
Logistic	0
Regression	2
By	2
26	0
Percentage	0
That	1

Determine the overall value of the word in the first document, the results at this stage are shown in Table 3.

TABLE 3 Result of Determining the Overall Value of the Word in the First Document

Words	Input_doc (A)	$(A)^2$
this	1	1
Paper	1	1
Showed	1	1
The	2	4
Machine	1	1
Learning	1	1
Algorithm	1	1
Svm	1	1
Is	1	1
Better	1	1
Than	1	1
Logistic	1	1
Regression	1	1
By	1	1
26	1	1
Percentage	1	1
That	1	1

The value obtain from $\sqrt{(A)^2} = 4.4721$

Determine the overall value of the word in the second document. The results at this stage are shown in Table 4

TABLE 4 Result of Determining the Overall Value of the Word in the Second Document

Words	Output_doc (B)	(B)2
this	3	9
Paper	1	1
Showed	1	0
The	16	32
Machine	6	36
Learning	4	16
Algorithm	0	0
Svm	0	0
Is	6	36
Better	0	0
Than	0	0
Logistic	0	0
Regression	2	4
By	2	4
26	0	0
Percentage	0	0
That	1	1

The value obtain from $\sqrt{(B)^2} = 19.0525$

Calculating Cosine Similarity. Using equation 1, the cosine similarity calculation results from 0.17458658, the cosine angle between 0 (zero) and 1 (one). To convert the percentage result, it must be multiplied by 100% so that the results of the similarity are 17.458658% with application result of 16.463398%.

B. Discussion

The proposed approach for plagiarism detection considers the values of vectors each word of entiredocument for the purpose of calculating the cosine similarity. The documents were taken from Kaggle, where itconsists of nearly 7000 rows of documents related to research papers. We have compared the result of cosinesimilarity with other similarity techniques like Jaccard similarity. It is observed that Cosine Similarity shows better results comparing the other one.

V. CONCLUSION

In this system, the cosine similarity based using tf-idfvectorizer website is deployed using python for detection of plagiarism. cosine similarity is a powerful technique for detecting plagiarism in documents. It is based on a mathematical model that can measure the similarity between two documents by comparing their corresponding vectors in a high-dimensional space. While it is not a definitive proof of plagiarism, it is an important tool that can help identify potential cases of plagiarism that require further investigation. The cosine similarity successfully detects the similar document.

REFFRENCE

1. D. Gunawan, C. A. Sembiring and M. A. Budiman, "The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents," IOP, 2018.

2. T. Foltýnek, D. Dlabolová, A. A. Naumeca, S. Razı, J. Kravjar, . L. Kamzola, J. G. Dib, Ö. Çelik and D. W. Wulff, "Testing of support tools forplagiarism detection," International Journal of Educational Technology in Higher Education, 2020.

3. A. Mishra and S. Vishwakarma, "Analysis of TF-IDF Model and its Variant for Document Retrieval," in *International Conference on Computational Intelligence and Communication Networks*, 2015, pp. 772-776,.

4. P. D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research*, pp. 141-148, 2010.

5. F. Mohammadi, "A New Approach To Focused Crawling: Combination of Text summarizing With Neural Networks and Vector Space Model," *ACSIJ Advances in Computer Science: an International Journal*, Vol. 2, Issue 3, No. 4, pp. 31-36, 2013.

6. S. Bhattacharjee, A. Das, U., Bhattacharya, S. K. Parui, and S. Roy, "Sentiment analysis using cosine similarity measure," in Proc. of IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), pp. 27-32, 2015.

7. L. H. Patil, and M. Atique, "A novel approach for feature selection method TF-IDF in document clustering," In Proc. of Advance Computing Conference (IACC), 2013 IEEE 3rd International, pp. 858-862, 2013.

8. P. Arabie, and G. De Soete, Clustering and classification, World Scientific, 1996.

.