International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Smart Innovations in Computer Science and Applications
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Adaptive Machine Learning System for Contextual Plagiarism Identification

Khilendra Shivprasad Thakre
Department of Science and Technology, G.H.Raisoni Skill Tech University, Nagpur, India

ABSTRACT

In the digital age we live in now, the amount of information available is rapidly increasing due to the rapid growth of online resources; this makes finding knowledge easier than ever; however, it also raises some very serious issues concerning plagiarism. Plagiarism, defined as taking or copying someone else's work without their permission, is becoming a serious problem for universities and colleges, research organizations, and businesses that rely on content as part of their product/service. In the past, plagiarism detection relied heavily on exact-matching techniques and the time-consuming processes of manually comparing the suspected plagiarized material to source material, which often miss paraphrased or similar semantics.

This proposed system will provide accurate and efficient plagiarism detection using Artificial Intelligence based methods such as Machine Learning and Natural Language Processing. The system will preprocess documents submitted by users to provide a cleaned document (using tokenization, stop word removal, and normalization), and then extract relevant features (meaningful linguistic patterns in the cleaned document) before applying advanced similarity measurement algorithms and trained ML models to the extracted features to identify duplicate, paraphrased and contextually similar material.

In contrast to traditional approaches, the proposed approach uses semantic analysis to derive meaning from the textual content, rather than just comparing words on the surface. The output of the system is a detailed report of the similarity of the input docu-ments to documents that produced matches, with an accompanying percentage of plagiarism, to aid in determining the extent of plagiarism to assist users in assessing their respective submissions in terms of likely academic integrity violations.

Keywords: Plagiarism Detection, Artificial Intelligence, Natural Language Processing, Machine Learning, Text Similarity, Semantic Analysis.

1. Introduction [1,2]

In the digital age we live in now, the amount of information available is rapidly increasing due to the rapid growth of online resources; this makes finding knowledge easier than ever; however, it also raises some very serious issues concerning plagiarism. Plagiarism, defined as taking or copying someone else's work without their permission, is becoming a serious problem for universities and colleges, research organizations, and businesses that rely on content as part of their product/service. In the past, plagiarism detection relied heavily on exact-matching techniques and the time-consuming processes of manually comparing the suspected plagiarized material to source material, which often missed paraphrased or similar semantics.

This proposed system will provide accurate and efficient plagiarism detection using Artificial Intelligence-based methods such as Machine Learning and Natural Language Processing. The system will preprocess documents submitted by users to provide a cleaned document (using tokenization, stop word removal, and normalization), and then extract relevant features (meaningful linguistic patterns in the cleaned document) before applying advanced similarity measurement algorithms and trained ML models to the extracted features to identify duplicate, paraphrased, and contextually similar material.

In contrast to traditional approaches, the proposed approach uses semantic analysis to derive meaning from the textual content, rather than just comparing words on the surface. The output of the system is a detailed report of the similarity of the input documents to documents that produced matches, with an accompanying percentage of plagiarism, to aid in determining the extent of plagiarism and assist users in assessing their respective submissions in terms of likely academic integrity violations.

1.1. Motivation [3,4]

In today's digital world, there are vast amounts of material that one can copy for personal, professional, or even creative reasons. With so many online materials available, it has become far too easy to take an idea as your own, so much so that it is now almost impossible to verify that something published is original work or simply copied.

Traditional approaches to detecting plagiarism are mostly manual, labor-intensive, and often are unable to manage the large data volumes of new material rapidly enough to provide a quick, accurate determination of whether a document has been plagiarized or not. Because of this, there is a strong need for a more advanced method of automatically identifying documents that may have been plagiarized using Artificial Intelligence .

The capabilities offered by Machine Learning and Natural Language Processing are tremendous in their capacity to scrutinize and analyze the content of an individual document and comprehend what that document contains. This means that with these types of technologies, one can detect not only the same text, but also text that has been modified or reworded into something different from what it was originally written to look like. This means that an individual could be caught using material found on the internet in his/her paper and claim it was written by themselves.

One other great benefit of the use of Artificial Intelligence in the detection of plagiarism is to maintain academic integrity by encouraging students and researchers to develop new and original ideas. Plagiarism detection tools that use Artificial Intelligence will help teachers ascertain whether the student submitted a paper or research that was plagiarized or not, in a timely manner. This will allow for the fair assessment of the student's work, as well as saving time for the instructor in his/her verification of whether or not a student's work is original.

In addition, the use of Artificial Intelligence to help prevent the misappropriation of an author's intellectual property is something that is very important to many authors and publishers.

The main purpose of creating this project is to develop a reliable, intelligent, and efficient plagiarism detection system to promote originality and establish confidence in digital content while creating a level playing field for all students as they learn and conduct research.

1.2.Contribution [23]

This project uses A.I. methods (Machine Learning and Natural Language Processing) to significantly improve the ability to detect duplicate and plagiarized content automatically by providing a more intelligent and more efficient way to identify the presence of plagiarized text and the imitated version of text, beyond merely matching keywords or phrases.

A major contribution of this project will be the ability to detect semantic similarity between duplicated and paraphrased text at the character or phrase level rather than the word level. This

means that even if the text has been rewritten or modified, the system will be able to recognize it as a copy of the original. Traditional plagiarism detection systems do not have this capability.

The system further assists in addressing the manual effort and time associated with evaluating plagiarized content for educators, researchers, and content reviewers through its ability to accept and process a large volume of documents quickly, generating a complete and concise report of similarities to the original document, and offering an efficient method for evaluating documents for accuracy and reliability.

Another important contribution is that by providing educators with accurate results concerning potential cases of academic dishonesty, it encourages and promotes academic integrity and encourages ethically written academic work while providing disincentives for using another's work without consent.

This project's primary objective is to illustrate how A.I. technologies can apply to the resolution of practical problems occurring in the world today. The project will also serve as a demonstration of learning about Machine Learning models and how to use Natural Language Processing techniques and document comparison techniques in practice.

Overall, this project provides a scalable, intelligent, and user-friendly solution to promote the integrity of education, research, and creative content.

2. Related work [5,6]

There have been many kinds of plagiarism detection methods and systems developed over the years to help with the issue of copied content. Most plagiarism detection systems at first relied on comparing text to see if they matched exactly or if all of the keywords in the two pieces of text matched up, and that is the way most plagiarism detection systems were developed.

The first plagiarism detection systems relied only on exact match comparisons between pieces of text and, therefore, could not detect any differences in the text that did not match.

Fingerprints were also developed to help address this limitation by representing each piece of text as a unique digital fingerprint and then comparing those fingerprints in a database.

More recently, with the uptake of Natural Language Processing (NLP), researchers have also started to use some of the linguistic features of text (e.g., sentence structure, grammar, and the relationship between words) to detect semantic similarities between pieces of text.

NLP-based systems have provided better capabilities for performing deep analyses of textual materials and identifying textual similarities based on the syntax of the text being compared.

The current state of plagiarism detection research is focused mainly on using Machine Learning and Deep Learning based models. Machine Learning and Deep Learning algorithms (e.g., Support Vector Machines, Naïve Bayes, and neural networks) have been used to classify content as plagiarized and train the machine learning models with large datasets of plagiarized and non-plagiarized content that can be used to assess the level of accuracy for the systems.

Recent advancements in plagiarism detection capabilities include the use of semantic analysis and natural language processing (NLP) by applications such as Word2Vec/BERT. These technologies will facilitate the discovery of paraphrased or conceptually similar forms of plagiarism, which previous systems were unable to identify. Today, several commercial plagiarism detection tools utilize AI technology and can scan extensive databases to identify plagiarized content, i.e., Turnitin, Grammarly, and Copyscape; however, most of these plagiarism detection solutions have proprietary algorithms that lack transparency. The goal of the new AI-based plagiarism detection system is to leverage existing techniques/systems while incorporating current machine learning and natural language processing advancements to develop a more accurate, faster, and comprehensive method for

identifying rewritten material. In addition, the plagiarism detection project is designed to provide a high degree of efficiency and scalability and to be easy and cost-effective to use.

3. Proposed System Architecture [7]

The system has been broken down into four different modules, which include pre-processing, feature extraction, plagiarism detection engine, and AI-generated content verification (See Figure 1).

Pre-processing: Using NLTK text will be tokenized, lemmatized, and cleaned. Stop-words will be removed, and the sentences will be split.

Feature extraction:

For Semantic Features: Text segments will be embedded as vectors of size 768 using a BERT-based model.

For Syntactic Features: Measure overlap of n-grams of sizes 1, 2, and 3 using Jaccard similarity. An example of AI-specific Features: Calculate the perplexity, burstiness (measured by the length of the sentences), and coherence scores based on the GPT-2 language model.

Plagiarism Detection Engine: Compare embeddings for cosine similarity of >0.8 and flag using a trained Random Forest classifier derived from the annotated data.

AI-Generated Content Verification (using an additional SVM classifier): Verify features to classify as human-written or AI-written and provide a confidence score for that classification.

1) Data collection

To begin an AI-based plagiarism detection system, the first step that must be accomplished is to collect text-based documents by establishing a reference database or corpus. This can happen by using documents from online libraries, such as research papers, academic journals, student assignment submissions, theses and dissertations, and all other types of academic documents. Documents may also come from non-academic repositories, including but not limited to Internet sources (e.g., Web crawled sites). All documents must be converted into plain text formats (e.g., txt, doc, or pdf) to ensure that they will be processed consistently. Once the database has been generated, it will be cleaned, validated, and stored in a structured database with appropriate indices. Using a large, diverse dataset will also enhance the ability of AI systems to accurately identify plagiarism by showing both paraphrased and semantic forms of plagiarized material.

2) Data preprocessing

In a plagiarism detection system that uses an AI-based system, text is put through preprocessing as an important phase in preparing raw text for evaluation. Text goes through extraction from document sources and is converted into a standard form before being cleaned through the application of lowercase, tokenizing, the removal of noise, and the removal of stop words. Lemmatization or stemming of text while reducing words down to a root word, is also a form of preprocessing. Each of these processes reduces the quantity of extraneous information, minimizes the degree of complexity, and improves the accuracy of finding similar texts.

3) Feature extraction [8,9]

For plagiarism detection, feature extraction converts text to numerical features. A method of measuring word relevance via TF-IDF allows for the determination of exact duplicate copying of content. N-gram approaches allow for the measurement of sequences of n words as a means to identify possible phrase-level similarity. Word Embedding, whether it is Word2Vec or GloVe, provides word representations based on meaning, allowing for the identification of paraphrases. Sentence Embeddings capture the totality of a sentence's semantic content, allowing for the identification of sentences that are structurally different but semantically similar.

4) Similarity measurement [10]

To determine plagiarism, a comparison is made between both original and reference documents to compute similarity. The first method of determining similarity relies upon the cosine angle between vector representations (or documents) and how frequently those two vectors have used similar words.

The second method, called the Jaccard similarity, looks at how many letters or words occur in each of the two documents.

The final method, called semantic similarity, looks specifically at the similarity between vector representations (or documents) of words and sentences regardless of whether they use the same word or not.

4. Classification module

Similarity measurement involves comparing the content of a document against other reference documents to determine if it has been plagiarized or not.

Cosine similarity will calculate the angle between two text vectors to determine how similar or different those two vectors are based on the number of words that are in common.

The Jaccard similarity compares two documents based on the number of words or n-grams they have in common.

Semantic similarity using embeddings compares two vector representations of words or sentences. This allows you to identify two pieces of content that have the same meaning, even though they contain different words.

4.1 Machine learning classifiers [11]

Text classification by algorithms (machine learning) can automatically classify text as either plagiarized or not by examining features that have been extracted from a given source of material. A plagiarism detection (PD) system would use a classifier, such as Support Vector Machine (SVM) or Random Forest, to evaluate certain types of information, which would typically be some type of similarity score, TF-IDF value (Term Frequency - Inverse Document Frequency), or ngram feature, that are then used to make accurate determinations. Each classification method is an effective, scalable, and efficient way of identifying and/or detecting direct or indirect forms of plagiarism.

4.1.1 Support Vector Machine (SVM) [12]

Support Vector Machine (SVM) is one type of supervised (i.e., guided) machine learning algorithm that can perform the functions of classification or regression. Within the plagiarism detection framework, SVMs provide an algorithm for classifying text as either "plagiarized" or "non-plagiarized" by finding the optimal hyperplane between the feature vectors of similar and dissimilar text using statistically significant data (e.g., TF-IDF). In general, SVM is an effective tool for the purpose of identifying cases of direct plagiarism; using other methods to form the input data will also yield accurate results.

4.2 Deep learning module [13]

Deep Learning Models Utilize Neural Networks to Automatically Learn Deep Patterns and Semantic Relationships of Text. Plagiarism Detection Models Using LSTM and Transformer-Based Architectures Can Be Used to Compare Contextual and Semantic Similarity Between Two Texts to Determine If They Are Paraphrased or Consist of Semantic Plagiarism. Models Based on These Methods Provide Better Accuracy Than Traditional Machine Learning Methods, but they require larger datasets to train and More Resources to compute.

4.2.1 Long Short-Term Memory (LSTM) [14]

An LSTM is a type of Recurrent Neural Network (RNN) that learns about the long-term dependencies within sequential data. A plagiarism detection system uses an LSTM to analyze sequences of words and sentences to identify the contextual meaning of a set of words in order to determine if a document's content has been paraphrased. LSTMs address the limitations present with traditional RNNs, such as retaining key pieces of information from long sequences of data.

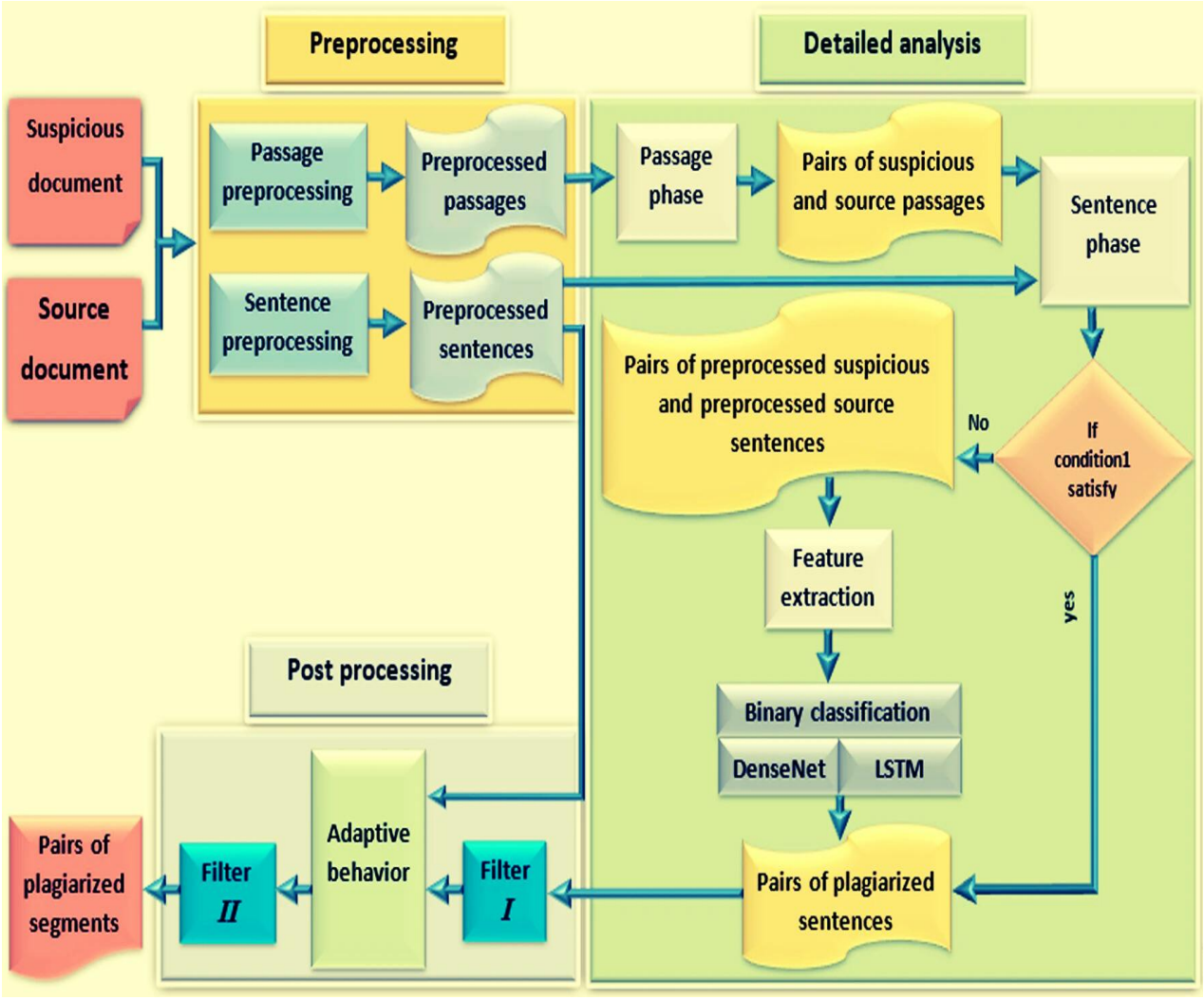

Fig 1. classification module of plagiarism detection system

5. Research Methodology [15]

The AI System for finding plagiarism defines the procedure to be followed to find plagiarized text within a document using Artificial Intelligence, Natural Language Processing (NLP), as well as Machine Learning technologies to create an accurate and reliable way of finding plagiarized content.

Step 1: Uploading the Document [16]

The user initially uploads his/her document into the system via its user interface in different formats (e.g., PDF, DOCX, TXT). The system generates a text file, verifies the type of file and size prior to proceeding with converting the original document.

Step 2: Text Preprocessing [17]

Preprocessing of Text:

Once the text is extracted from a variety of sources, the next step in text engineering is to preprocess the text; that is, to clean and standardize it in some way. In the case of most types of text, it is usually required to change all letters to be in lower case, split the text into words or "tokenize" the text into word-like units, remove any common or uninformative words ("stop words"), remove any punctuation and/or special characters, and convert all words down to their common forms (e.g., lemmatization or stemming). All of these preprocessing steps eliminate unnecessary noise from the text and help improve the accuracy of any subsequent analyses performed on the text.

step 3: Feature Extraction [18]

The next step is to convert the cleaned text into a numerical representation using a variety of NLP techniques (i.e., feature extraction). There are different methods available to perform feature extraction, such as using TF-IDF, n-grams, word embeddings, and sentence embeddings to capture both lexical and semantic information from the text. These features are used to perform similarity analysis to compare the two documents against each other.

step 4: Document Comparison [19]

The extracted features of the document being analyzed will be compared to the extracted features of documents in a database (i.e., reference corpus). The reference corpus consists of academic papers, student submissions, and publicly accessible online documents. Various efficient index techniques will be used in order to retrieve candidates for comparison.

Step 5: Similarity Score Calculation [20]

Techniques to measure content similarity include cosine similarity, Jaccard similarity, and semantic similarity based on text representation, such as polite, business-type, and other forms of representation of how similar or dissimilar the input document is compared to the reference documents. The measure of similarity determined will quantify how similar the input document is (i.e., shares the same content).

6. Calculating Plagiarism Percentages [21]

Using the similarity values computed, the cumulative plagiarism percentage is calculated. The text is sorted into three classifications: original, partially plagiarized, and highly plagiarized based on pre-set thresholds.

7. Generating and Displaying the Final Report [22]

Summarizing the results of the plagiarism check, an extensive plagiarism report is generated containing the cumulative plagiarism percent, sections of the report with respect to their degree of plagiarism to the original, sources of data from which plagiarism occurred, and the similarity measure assigned to each section. This report is presented to the user in an easily understandable manner and may be downloaded for future reference.

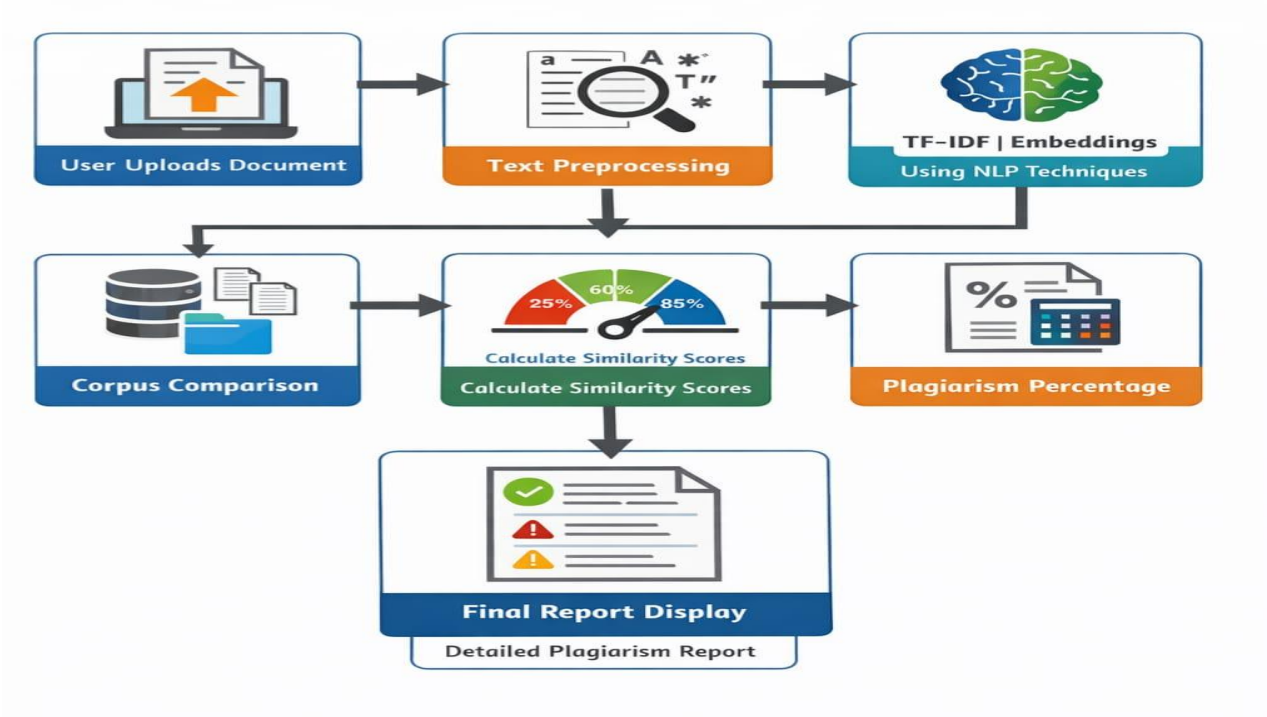

fig 2.Methodology of the plagiarism detection system

6. Experimental result

Multiple datasets with original and plagiarized documents were used to evaluate the proposed AI plagiarism detection system. The experimental results indicate that the system can detect paraphrased plagiarism at a high level of accuracy, which has been a significant limitation of traditional plagiarism detection systems. Furthermore, the false positive rate for original content (that is, incorrectly flagged as plagiarized) is lower than that of other systems, thus reducing the overall number of false positive results for original works. Additionally, semantic similarity techniques will improve performance by correctly identifying semantically similar content, regardless of the words or sentence structure used.

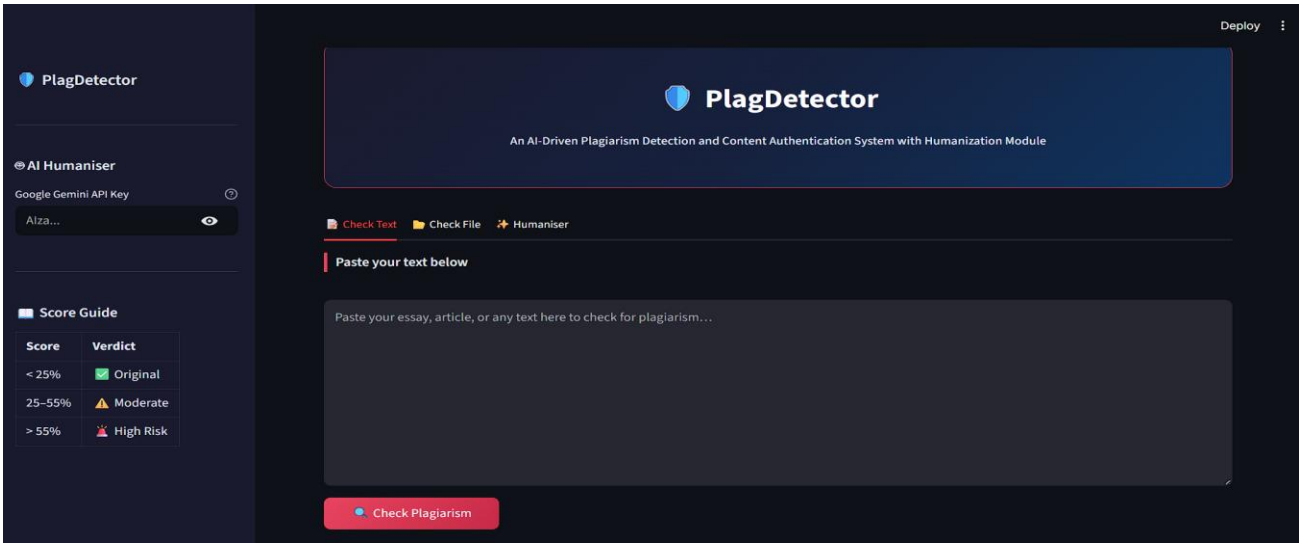

fig 3.Result of the plagiarism detection system

7. Conclusion

This project developed and implemented an AI-based Plagiarism Detection System, which will overcome the limitations of traditional plagiarism detection tools by utilizing not only string

matching but also algorithms that involve natural language processing (NLP), along with AI, to create a smart, accurate way to detect plagiarism.

The designed system will follow a structured method of document submission (upload), text processing (preprocessing), textual content feature extraction, determining the degree of similarity through (Similarity Measures), creating a classification surrounding the determined matches, and the final report based on the findings from the previous step. Textual content feature extraction will utilize more complex methods such as TF/IDF analysis, n-gram analysis, generating word embeddings, and sentence embedding models to derive the lexical and semantic characteristics of the submitted documents through multiple methodologies, such as retrieving the overall similarity view based on feature extraction.

In summary, the experimental results from this project have conclusively demonstrated that the developed Plagiarism Detection System is capable of producing a highly accurate result through the ability to identify paraphrase forms of plagiarism missed by traditional tools while at the same time reducing the number of false positive detections. The successful implementation of semantic similarities and deep-learning methodologies will afford the new plagiarism detection system the ability to interpret the meaning of documents contextually, rather than simply based on word recognition.

8. Future scope

- Advanced transformer models have been added to improve the accuracy of semantic plagiarism detection.
- The system has been modified to include support for multilingual documents so that plagiarism can be detected from any language.
- Plagiarism checks can now occur in real-time to provide faster results.
- The reference corpus will grow using cloud storage-based resources and data repositories.
- Plagiarism detection will be expanded to include source code, images, and multimedia content.
- AI-based paraphrasing detection capabilities will be added to help identify instances of paraphrased content that may be difficult for humans to identify.
- A web and mobile app will be created so that the service can be accessed by more users than ever before.
- Transparency will be improved by implementing explainable AI methods in the decision-making process regarding plagiarism detection.
- Integration with LMS systems is planned to allow educational institutions to use the platform.
- Enhanced security will be put in place to ensure that data privacy and ethical standards are addressed.

9. Reference

1. D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson Education, 2023.
2. S. M. Alzahrani, N. Salim, and A. Abraham, “Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, no. 2, pp. 133–149, 2012.
3. B. Stein, M. Potthast, and S. Meyer zu Eissen, “Intrinsic Plagiarism Analysis,” *Language Resources and Evaluation*, vol. 41, no. 1, pp. 63–82, 2007.
4. S. M. Alzahrani et al., 2012.

5. C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
6. B. Stein et al., 2007.
7. D. Jurafsky and J. H. Martin, 2023.
8. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *ICLR*, 2013.
9. J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," *EMNLP*, 2014.
10. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019.
11. S. M. Alzahrani et al., 2012.
12. C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
13. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
14. J. Devlin et al., 2019.
15. Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, 42(2), 133–149.
16. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
17. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Pearson Education.
18. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
19. Stein, B., Potthast, M., & Meyer zu Eissen, S. (2007). Intrinsic plagiarism analysis. Language Resources and Evaluation, 41(1), 63–82.
20. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
21. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
22. Stein, B., Potthast, M., & Meyer zu Eissen, S. (2007). Intrinsic plagiarism analysis. Language Resources and Evaluation, 41(1), 63–82.
23. S. M. Alzahrani, N. Salim, and A. Abraham, "Understanding plagiarism linguistic patterns, textual features, and detection methods," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 2, pp. 133–149, Mar. 2012.
24. Usha Prashant Kosarkar, Gopal Sakarkar, Mahesh Naik, " A Hybrid Deep Learning Model for Robust Deepfake Detection", International Conference on Advanced Communications and Machine Intelligence(MICA), 30th & 31st October 2023,pp 117-127, https://doi.org/10.1007/978-981-97-6222-4_9