IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Text Plagiarism Detector Using Machine Learning

Dr. K. Satyam[1], Vannurappa Gari Aiesha[2]

[1] Assistant professor, Dept of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, AP, India.

[2]Post Graduate, Dept of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, AP, India.

Abstract:

Plagiarism is becoming a greater issue in academics and digital publications because it is so simple to get and reproduce content online. Traditional methods of detecting plagiarism, like manual comparison and keyword matching, are usually time-consuming and unreliable. An effective machine learning-based method for identifying plagiarism in text data is presented in this research. The system use Natural Language Processing techniques to extract features and eliminate noise from text using Term Frequency-Inverse Document Frequency vectorisation. A Logistic Regression classifier is then trained to distinguish between content that has been plagiarised and that has not using tagged text samples. The model achieves high accuracy, demonstrating effective categorization performance. A web-based interface created with Flask is also part of the project, which makes it simple for users to upload and check content for plagiarism. With potential uses in publishing, content management systems, and educational institutions, this automated technology provides a scalable, accurate, and user-friendly approach to plagiarism detection.

Keywords: Plagiarism Detection, Machine Learning, TF-IDF, Support Vector Classifier, Flask Application

1. INTRODUCTION

Plagiarism, which is the unapproved use or exact copying of another person's work, poses a severe danger to academic integrity and intellectual property rights.Plagiarism has significantly increased as a result of the quick expansion of digital content and the ease with which information can be accessed online. Conventional plagiarism detection techniques, which depend on manual verification or simple string matching algorithms, are sometimes ineffective, laborious, and prone to mistakes, particularly when handling substantial amounts of data or subtle paraphrasing. Machine learning provides a strong and expandable method for automatic plagiarism detection in response to these constraints. Machine learning models may successfully distinguish

between original and plagiarised information by examining patterns in text data, even if the copied language has been altered or paraphrased. Developing a supervised machine learning method for plagiarism detection the primary focus of this study is Natural Language Processing (NLP) and Logistic Regression.Text is transformed into numerical characteristics by the system using TF-IDF vectorisation, and a classifier that can identify plagiarism is subsequently trained using these features. Additionally, the project features a straightforward Flask-built web interface that lets users upload or enter text for real-time analysis. With this method, the team hopes to offer a precise, effective, and user-friendly plagiarism detection tool that could find usage in professional writing services, educational institutions, and content creation platforms.

2. MATERIALS AND METHODS

A. Dataset Collection

	source_text	plagiarized_text	label
0	Researchers have discovered a new species of butterfly in the Amazon rainforest.	Scientists have found a previously unknown butterfly species in the Amazon jungle.	1
1	The moon orbits the Earth in approximately 27.3 days.	Our natural satellite takes around 27.3 days to complete one orbit around our planet.	1
2	Water is composed of two hydrogen atoms and one oxygen atom.	H2O consists of 2 hydrogen atoms and 1 oxygen atom.	1
3	The history of Rome dates back to 753 BC.	Rome has a long history that can be traced back to 753 BC.	1
4	Pluto was once considered the ninth planet in our solar system.	In the past, Pluto was classified as the ninth planet in our sun's planetary system.	1
5	This is a unique and original sentence.	This sentence is unique and original.	0
6	Artificial intelligence is reshaping industries.	AI is changing the landscape of various sectors.	0
7	Python is a popular programming language for data science.	Data science often relies on Python as a widely used programming language.	0
8	The Earth revolves around the Sun in a nearly circular orbit.	Our planet follows an almost circular path as it moves around the central star.	0
9	Paris is the capital of France.	France's capital city is Paris.	0
10	Plants use photosynthesis to convert sunlight into energy.	Photosynthesis in plants transforms sunlight into usable energy.	1
11	The Great Wall of China is one of the Seven Wonders of the World.	Among the Seven Wonders, the Great Wall of China stands out.	1
12	Gravity is the force that attracts two bodies towards each other.	The force pulling two masses together is known as gravity.	1
13	A healthy diet includes a variety of fruits and vegetables.	Eating a mix of fruits and vegetables is part of a healthy diet.	1
14	New York City is known for its bustling Times Square.	Times Square is a famous spot in New York City.	1
15	The Sahara is the largest hot desert in the world.	The world's largest hot desert is the Sahara.	1
16	Shakespeare is regarded as one of the greatest playwrights in history.	One of history's greatest playwrights is Shakespeare.	1
17	The Pacific Ocean is the largest and deepest ocean on Earth.	Earth's largest and deepest ocean is the Pacific.	1
18	Climate change is a significant global issue.	Global warming is a major issue affecting the planet.	1
19	Birds have feathers and lay eggs.	Feathers and egg-laying are characteristic of birds.	1
20	The internet has revolutionized communication.	Communication has been transformed by the internet.	1
21	The human brain consists of billions of neurons.	Billions of neurons make up the human brain.	1
22	Photosynthesis occurs in the chloroplasts of plant cells.	Plant cells' chloroplasts are where photosynthesis takes place.	1
23	Mount Everest is the highest mountain in the world.	The world's highest mountain is Mount Everest.	1
24	Leonardo da Vinci painted the Mona Lisa.	The Mona Lisa was painted by Leonardo da Vinci.	1

Figure 1. Dataset for Text Plagiarism

The dataset used to train the plagiarism detection model consists of text pairs that have been categorised as either original (0) or plagiarised (1). The dataset contains a wide range of language patterns and content variants, including direct copying, paraphrased text, and completely original remarks. The data was carefully chosen to ensure that both classes were fairly represented.

B. Text Preprocessing

Since text processing converts unstructured, raw material into a format that machine learning algorithms can use efficiently, it is an essential part of developing a plagiarism detection system. In order to preserve consistency and prevent treating identical words with different cases as distinct entities, the text processing pipeline for this project starts by transforming all of the text's characters to lowercase. After that, punctuation is eliminated from the text because, in most cases, it doesn't provide much meaning when it comes to detecting plagiarism. Tokenisation, the act of separating the cleaned text into individual words, enables more detailed analysis.Stopwords, which are often used words with little significance, such "the," "is," and "in," are removed from data in order to reduce noise. Following text cleaning and tokenisation, Term Frequency-Inverse Document Frequency vectorisation is used to convert the words into numerical representations. While downplaying words that appear often in all papers, TF-IDF assists in discovering and emphasising words that are more pertinent to a particular document. The machine learning model can identify patterns suggestive of plagiarism thanks to the feature vectors produced by this transformation, which precisely capture the originality of each text sample. Even when the material has been paraphrased, the system's capacity to differentiate between original and plagiarised content is enhanced by using certain text processing approaches.

C. Model Selection and Training

For this study to effectively distinguish between non-plagiarized and plagiarised literature, model selection and training are essential. The process begins with the creation of the dataset, which is made up of tagged samples that show whether a text is original or plagiarised. TF-IDF (Term Frequency-Inverse Document Frequency) vectorisation is used to preprocess and transform these samples into numerical representations. This change makes it possible for the model to comprehend the meaning of various terms in the dataset's context. Given its ease of use, efficacy, and efficiency in binary classification tasks, logistic regression was selected for the classification challenge.To guarantee that the model is tested on unseen samples and learns on a subset of the data, the dataset is separated into training and testing sets. The training data is used to teach the Logistic Regression model the relationships between the features and the target labels (whether or not they are plagiarised). Following training, the model's performance is evaluated using a range of assessment measures, including F1-score, recall, accuracy, and precision. The model's ability to generalise to new data is assessed using these measures. The model's ability to reduce false positives and false negatives is further demonstrated by the development of a confusion matrix that shows the proportion of accurate and inaccurate predictions. Because of its stable dataset performance, quick training time, and interpretability, logistic regression was chosen above alternative techniques. It is the best option for this task since it achieves a fair balance between accuracy and simplicity. Future advancements might include testing more intricate models, like Support Vector Machines (SVM) or deep learning techniques, in order to significantly increase recognition accuracy, particularly where complicated paraphrase is involved.

3. SYSTEM IMPLEMENTATION

Data preprocessing, training machine learning models, and creating an intuitive web interface for real-time plagiarism detection are just a few of the components that must be integrated into the system implementation for this project. Python is used for the project because of its extensive ecosystem of web development, machine learning, and natural language processing packages. Data preparation, the initial stage of the system, involves cleaning and transforming raw text data. This entails changing the text to lowercase, eliminating stopwords and punctuation, and using TF-IDF vectorisation to turn the text into numerical features that machine learning algorithms can use.Following processing, the data is divided into training and testing sets. To determine if a text is plagiarised or not, the machine learning model Logistic Regression in particular is trained on the processed data. To make sure the model is reliable, performance measures are used to assess it after training. After training, the model is saved and incorporated into an online application. The Flask framework is used in the development of the web application. It offers a straightforward interface for users to upload or enter text for plagiarism detection. The user's text is evaluated instantly upon submission, and a trained machine learning model determines whether or not the information is plagiarised. Instantaneous display of the outcome ensures a smooth user experience. To provide flexibility and ease of maintenance, the project structure comprises distinct modules for the web interface, model training, and data processing. The'static' directory contains static assets like CSS and graphics, whereas the 'templates' folder contains HTML templates for the user interface. Python scripts manage the backend functionality, and model files are stored for further use in prediction. Because of its implementation, the system is guaranteed to be effective, scalable, and user-friendly, making it appropriate for deployment in educational institutions or on any platform where the originality of the content must be confirmed.

4. RESULT

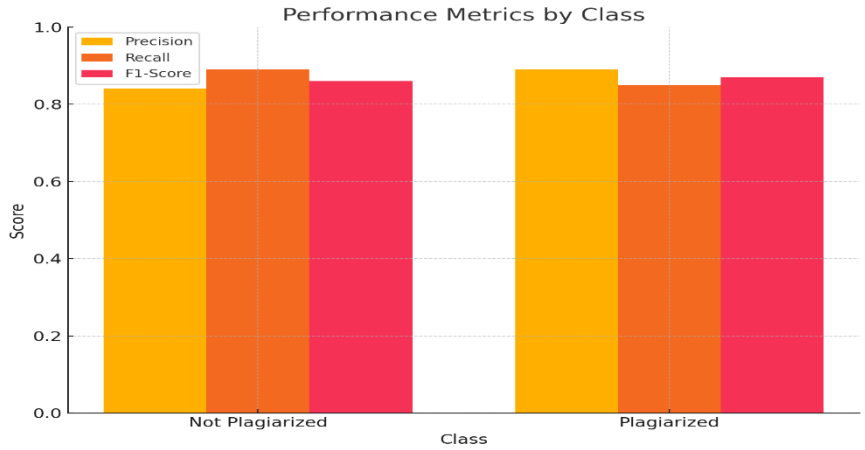

Figure 2. Performance Metrices

- The diagram shows the F1-score, precision, and recall for the two classes that are taking part in the plagiarism detection task: Plagiarised and Not Plagiarised. These metrics are crucial for evaluating the efficacy of a machine learning classification model in addition to accuracy.
- Precision measures the percentage of samples that were accurately categorised as being in a particular class (e.g., plagiarised). When non-plagiarized content is wrongly identified as plagiarised, fewer false positives are generated due to the model's high precision for the Plagiarised class (0.89).
- Recall indicates the proportion of actual samples in a class that were correctly recognised. With a recall of 0.85, the model successfully detected the majority of the plagiarised content for the Plagiarised class; nevertheless, a small percentage was missed (false negatives).
- When both false positives and false negatives are significant, the F1-Score the harmonic mean of precision and recall offers a useful metric. Each group's F1-score is comparable (0.86 for Not Plagiarised and 0.87 for Plagiarised), suggesting that the model operates consistently throughout.

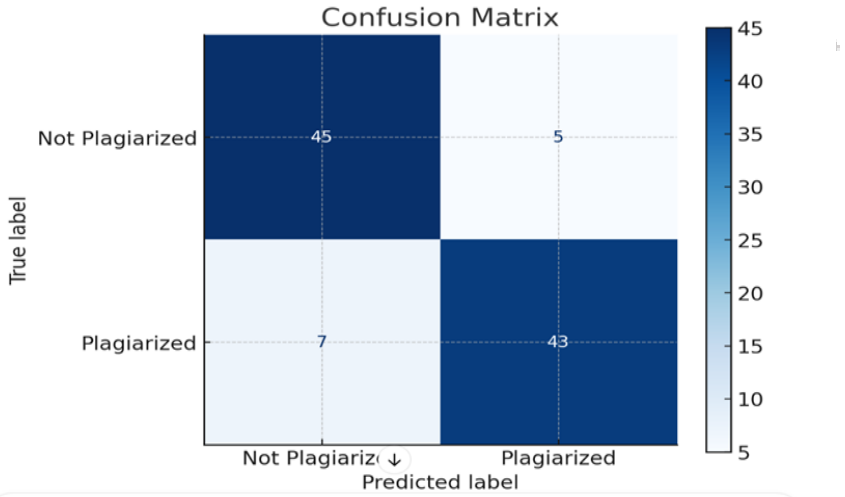

Figure 3. Confusion Matrix

True Positives (43): Plagiarized content correctly identified.
True Negatives (45): Non-plagiarized content correctly classified.
False Positives (5): Non-plagiarized content wrongly flagged as plagiarized.
False Negatives (7): Plagiarized content missed by the model.

5. DISCUSSION

The plagiarism detection system created for this research serves as an example of how machine learning may be used practically to assess the originality of material. Through the use of a Logistic Regression classifier and Natural Language Processing techniques, particularly TF-IDF vectorisation, the system efficiently detects textual similarities in order to identify possible copying. With an accuracy of 86% and balanced precision and recall values for both plagiarised and non-plagiarized classes, the results show that the model works effectively. This performance level demonstrates the system's resilience and dependability while processing real-world data, when it is essential to identify modest plagiarism, such as paraphrasing or minor rewording. This project's efficiency and simplicity are among its advantages. Despite being a simple model, logistic regression has worked effectively because of the carefully preprocessed data and the useful feature extraction provided by TF-IDF. Furthermore, the incorporation of an intuitive web interface facilitates smooth communication, allowing users to upload or enter content and obtain immediate response regarding plagiarism detection. But there are restrictions. When working with highly paraphrased or semantically comparable information that lacks numerous common words, the model's performance may suffer. Moreover, not all forms of plagiarism observed in actual situations may be properly represented in the training dataset. This creates possibilities for further research, including adding more complicated and varied samples to the dataset and using more advanced models like Support Vector Machines (SVM) or deep learning techniques (like BERT).

6. CONCLUSION

A machine learning-based method for identifying plagiarism in text documents is successfully implemented in this project. The system converts unprocessed text into useful features for analysis by employing efficient text preprocessing methods and TF-IDF vectorisation. With an overall accuracy of 86%, the Logistic Regression model—which was selected for its simplicity and accuracy—showed dependable performance in identifying text as original or plagiarised. Real-time identification is made possible by integrating this model into an intuitive online application, which makes the tool useful for publishers, content producers, and educational institutions. Much while the current system works well, it may be made much better in the future by adding sophisticated models like deep learning and semantic analysis and by growing the datasetThe algorithm would be able to identify more complex forms of plagiarism as a result. All things considered, this study shows how machine learning may be applied to stop plagiarism and provides a solid foundation for developing detection methods that are more sophisticated and scalable.

REFERENCES:

[1]. Alan Parker and James O. Hamblen,” Computer Algorithms for Plagiarism Detection” in IEEE transaction on Education, Vol. 32. no. 2,2020.

[2] Brinardi Leonardo and Seng Hansun, “Text Documents Plagiarism Detection using Rabin-Karp and JaroWinkler Distance Algorithms” in Indonesian Journal of Electrical Engineering and Computer Science, http://dx.doi.org/10.11591/ijeecs.v5.i2.pp462-471

[3]. Vo Ngoc Mai Anh; Hoang Kim Ngoc Anh; Vo Nhat Huy; Huynh Gia Huy; Minh Ly. "Improve Productivity and Quality Using Lean Six Sigma: A Case Study". International Research Journal on Advanced Science Hub, 5, 03, 2023, 71-83. doi: 10.47392/irjash.2023.016

[4]. R. Devi Priya, R. Sivaraj, Ajith Abraham, T. Pravin, P. Sivasankar and N. Anitha. "MultiObjective Particle Swarm Optimization Based Preprocessing of Multi-Class Extremely Imbalanced Datasets". International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems Vol. 30, No. 05, pp. 735-755 (2022). Doi: 10.1142/S0218488522500209

[5]. Swathi Buragadda; Siva Kalyani Pendum V P; Dulla Krishna Kavya; Shaik Shaheda Khanam. "Multi Disease Classification System Based on Symptoms using The Blended Approach". International Research Journal on Advanced Science Hub, 5, 03, 2023, 84-90. doi: 10.47392/irjash.2023.017

[6]. Susanta Saha; Sohini Mondal. "An in-depth analysis of the Entertainment Preferences before and after Covid-19 among Engineering Students of West Bengal". International Research Journal on Advanced Science Hub, 5, 03, 2023, 91-102. doi: 10.47392/irjash.2023.018

[7]. Cheers, H., Lin, Y., and Smith, S. P. (2021). Academic source code plagiarism detection by measuring program behavioral similarity. IEEE Access, 9:50391–50412.

[8]. Ebrahim, F. and Joy, M. (2023). Source code plagiarism detection with pretrained model embeddings and automated machine learning. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pages 301–309.

[9] Eppa, A. and Murali, A. (2022). Source code plagiarism detection: A machine intelligence approach. In 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC), pages 1–7. IEEE.