


A celestial scene featuring a planet and a moon against a starry background. The planet is on the left, partially illuminated, with a bright light source behind it creating a lens flare. The moon is to its right, also partially illuminated. The background is dark with many small stars.

AI Plagiarism Detector

This presentation discusses an AI-based plagiarism detection system utilizing the MIT Plagiarism Detector dataset. It explores various machine learning models such as logistic regression, SVM, naive bayes, and random forest for predicting instances of plagiarism.



Introduction to Plagiarism Detection

Detection is vital for maintaining quality in educational and professional settings.

1 Importance of Plagiarism Detection

Plagiarism compromises academic integrity and originality.

2 Overview of AI in Detection

AI techniques enhance accuracy and efficiency in detecting plagiarism.

3 Dataset Utilization

The MIT dataset aids in training various machine learning models.

Machine Learning Models



Logistic Regression

Predicts binary outcomes, distinguishing original and plagiarized texts.



Support Vector Machine

Finds the optimal hyperplane for data separation in high-dimensional spaces.



Naive Bayes Classifier

Probabilistic model efficient for large datasets, commonly used in text classification.



Random Forest Algorithm

Ensemble method that reduces overfitting by building multiple decision trees.

Data Preprocessing Steps

Raw text data undergoes cleaning to remove noise, symbols, and irrelevant content, which aids in improving model performance. Techniques such as TF-IDF are applied to convert textual data into numerical vectors, representing text effectively for model input. Instances are classified into 'plagiarized' or 'original' to prepare labels for supervised learning, establishing the foundation for the training process.



Data Cleaning

Remove noise and irrelevant content.



Feature Extraction

Convert text to numerical vectors using TF-IDF.



Label Encoding

Classify into 'plagiarized' or 'original'.



Model Training and Evaluation

Training Process

Models are trained on dataset portions to optimize accuracy.

Validation Techniques

Cross-validation assesses model performance for reliability.

Metric Analysis

Metrics like accuracy help compare model effectiveness.



Results and Findings

Model Comparison

Random forest and SVM outperform naive bayes.

Error Analysis

Misclassified instances reveal common patterns.

Textual Patterns Insight

Key linguistic features in plagiarized texts.



Future Work

Enhancing Dataset

Expanding the dataset with diverse samples.

Exploring Advanced Techniques

Incorporating LSTM models for better context.

Real-world Application

Deployment for automatic plagiarism checks.



Conclusion

1

Summary of Findings

AI shows potential in detecting plagiarism.

2

Significance of Continued Research

Research is vital to adapt to evolving text manipulation.

3

Emphasis on Ethics

Ethical implications must be considered in AI tools.