



School of Computer Science Engineering and Information Systems (SCORE)

Fall Semester – December 2025 (MIS)

Course Name: Natural Language Processing

FACULTY: Prof. Dr. SenthilKumar N

Course Code: SWE1017

Course Slot: E2

J Component Final Report

PROJECT TITLE: DOCUMENT SIMILARITY CHECKER
FOR ACADEMIC WRITING

Student Details

S. No.	Register Number	Student Name
1	22MIS0361	BHAVA DHARINI J
2	22MIS0375	KAVIYA G

TABLE OF CONTENTS

S. NO.	TITLE	PAGE NO.
1	Abstract	3
2	Introduction	3
3	Problem Statement	4
4	Literature Survey	5
5	Dataset Description	14
6	Unique Aspects	15
7	Architecture Diagram	16
8	Proposed Model / Approach	17
9	Implementation	19
10	Evaluation Metrics	34
11	Observation	35
12	Tools (Framework Environment)	35
13	Conclusion	37
14	Future Work	37
15	Evidence of Collaboration	38
16	References	39

1. ABSTRACT

Plagiarism detection in academic writing requires more than just matching identical words — it must also recognize rephrased or semantically similar content. This project presents a Document Similarity Checker that combines both traditional and deep learning approaches for effective plagiarism detection. The system first uses TF – IDF with Cosine Similarity to identify direct or word – level similarities, ensuring fast comparison. Then, it employs fine – tuned transformer – based models – BERT – base – NLI – mean – tokens , all – MiniLM - L6 - v2, and all – mpnet – base - v2 (SBERT variants) – to capture deeper, meaning-level relationships between documents. The models were trained and evaluated on the MIT Plagiarism Dataset, enabling them to understand real academic writing styles. Experimental results show that the hybrid method achieves high accuracy and reliability, effectively detecting both exact and paraphrased plagiarism in large documents. This approach bridges the gap between speed and semantic understanding, offering a practical solution for academic plagiarism detection.

2. INTRODUCTION

In today's academic world, plagiarism has become a growing problem as students and researchers have easy access to large amounts of online information. Many people unintentionally or intentionally reuse others' work, making it difficult for instructors to check originality. Traditional plagiarism checkers often depend on exact word matching and fail to detect rephrased or paraphrased content. Therefore, there is a strong need for systems that can understand the meaning behind the text and not just the words used. This project “Document Similarity Checker for Academic Writing”, aims to build an intelligent plagiarism detection system that goes beyond simple keyword matching. The motivation for this work is to help educators and institutions identify both direct copying and semantic plagiarism more accurately.

To achieve this, the system combines both traditional NLP and modern deep learning methods. The traditional method, TF – IDF with Cosine Similarity, helps to quickly detect direct overlaps, while Transformer – based models like BERT and Sentence – BERT (SBERT) identifies deeper meaning – level similarities between sentences and documents. By fine – tuning these models on academic

datasets such as the MIT Plagiarism Dataset, the system learns to recognize how ideas can be expressed differently but still convey the same meaning. Overall, this project demonstrates how combining statistical and deep learning approaches can create a more accurate, efficient, and reliable plagiarism detection system for real academic writing.

3. PROBLEM STATEMENT

This project focuses on solving the challenging NLP problem of detecting semantic plagiarism – identifying when two academic documents express the same ideas using different words. Traditional plagiarism detection tools mainly depend on exact word or phrase matching, which fails when sentences are paraphrased or restructured. To address this limitation, our project aims to detect both direct and meaning – level similarity between academic texts.

The system combines three types of models, each contributing unique strengths:

- TF-IDF with Cosine Similarity detects exact and near – exact text overlaps efficiently.
- BERT – base – NLI – mean – tokens, a transformer – based model, captures context and sentence meaning beyond simple keywords.
- SBERT variants — all – MiniLM - L6 - v2 and all – mpnet – base - v2 – provide lightweight and high – performing embeddings, enabling faster yet accurate semantic comparisons.

4. LITERATURE SURVEY

S. No.	Author & Year	Title	Methodology / Tools	Key Findings	Limitations
1	Nicholas Gahman, Vinayak Elangovan (2023)	<i>A Comparison of Document Similarity Algorithms</i>	Categorized algorithms into three types: (1) Statistical (SBERT + Cosine Similarity), (2) Neural Network – based (MT – DNN, XLNet), and (3) Knowledge / Corpus – based (Lin Measure + String Similarity). Evaluated on MRPC, AFS, SICK – R, SICK – E datasets using Pearson, Spearman, and Accuracy metrics.	The MT – DNN neural model performed best across all benchmarks. SBERT + Cosine and Lin Measure performed moderately well. XLNet had inconsistent results, especially in regression tasks.	XLNet's instability and high GPU usage were drawbacks. Statistical and corpus – based methods required complex preprocessing. No single algorithm was best for all domains.
2	Rohitha Ravinder (2023)	<i>A Comparison of Vector – based Approaches for Document Similarity Using the</i>	Used the RELISH biomedical corpus from PubMed. Compared five techniques: Word2Vec (word2doc), Doc2Vec, Dictionary – based NER + TF – IDF, Hybrid Doc2Vec + NER, and BERT – based models (BioBERT, SciBERT). Evaluated with nDCG and precision metrics.	The Word2Vec approach gave the best performance overall. Among BERT – based models, SciBERT performed strongest. NER + TF – IDF methods were less accurate.	Focused only on biomedical domain, so results may not generalize. RELISH dataset biases may affect outcomes. Embedding models required high computational resources.

		<i>RELISH Corpus</i>			
3	S. V. Moravvej (2023)	<i>A Novel Plagiarism Detection Approach Combining BERT – based Embedding, Attention – LSTMs & Differential Evolution Algorithm</i>	Used BERT embeddings for sentence pairs, attention – based Bi – LSTM for similarity prediction, and an Improved Differential Evolution algorithm to optimize weights. Used Focal Loss to handle class imbalance. Evaluated on MSRP, SNLI.	Achieved strong performance (Accuracy \approx 88%, F – score \approx 0.93). Combining BERT + Attention + LSTM + DE optimization significantly improved plagiarism detection over traditional lexical methods.	High computational cost due to BERT's large size. Class imbalance partially handled. Real – world performance across languages or styles still uncertain.
4	Jiarong Xian (2024)	<i>BERT – Enhanced Retrieval Tool for Homework Plagiarism Detection System</i>	Created a synthetic dataset (\approx 33K text pairs) using GPT – 3.5. Extracted embeddings with SBERT, used FAISS for fast retrieval, and applied an MLP classifier. Evaluated using accuracy, precision, recall, and F1 score.	Achieved \approx 98.9% accuracy and F1 \approx 0.99. FAISS enabled fast and scalable retrieval. Covered multiple plagiarism types (exact copy, imitation, paraphrase). Built an interactive demo tool.	Dataset was synthetic, not real student data – may not reflect real – world plagiarism. Struggles with subtle or deeply paraphrased text. Doesn't handle multilingual or code plagiarism.

5	Kanak Kalyani (2023)	<i>PlagCheck: An Efficient Way to Identify Plagiarism Using BERT</i>	Used BERT embeddings with Cosine similarity to compare essays and assignments. Tested multiple embedding variants for academic writing.	The BERT – based cosine method gave high accuracy and efficiency in plagiarism detection for student essays.	Limited to text – only English content (no code/multilingual data). May miss heavily paraphrased or translated plagiarism.
6	Deblina Mazumder Setu (2025)	<i>A Comprehensive Strategy for Identifying Plagiarism in Academic Submissions</i>	Three – phase hybrid pipeline: (1) TF – IDF + Cosine for direct copy, (2) BERT for semantic paraphrase, (3) Web scraping + TF – IDF for online sources. Evaluated on MIT Plagiarism Dataset (367K samples).	Achieved ~71% accuracy, ~80% recall, ~74% F1. Detected both direct and paraphrased plagiarism effectively. Useful for academic submissions.	Moderate accuracy (false positives & negatives). BERT + web – scraping is resource – intensive. May miss cleverly rewritten or non – textual plagiarism.
7	Kartik Vyas (2023)	<i>Semantic Similarity and Plagiarism Checker</i>	Combined Cosine Similarity, n – grams, Word2Vec, BERT, and classical methods (Jaccard, Rabin – Karp). Used OCR for handwritten text. Evaluated on MSRPC and SICK datasets.	Achieved 78.2% accuracy (binary classification) and $MSE \approx 0.30$ (semantic regression). Integrated both semantic and lexical measures effectively.	OCR accuracy for handwriting only ~70%. False positives due to thresholding. High compute cost for BERT – based models.
8	K. Srinivasa	<i>An Interactive Plagiarism Checker with</i>	Built a GUI – based tool with two stages: (1) TF – IDF + Cosine for broad detection, (2) Levenshtein	The tool provided visual highlighting and improved accuracy for large	Edit-distance methods limited for paraphrasing.

	Babu (2024)	<i>Text Processing and Similarity Analysis</i>	Distance (edit distance) for fine matching and highlighting.	documents. Easier to use for students / instructors.	Scalability issues for very large document sets.
9	Abdul Wahab Qurashi (2020)	<i>Document Processing Methods for Semantic Text Similarity Analysis</i>	Used Word2Vec embeddings for multi – word sentences. Compared Jaccard vs Cosine similarity in railway safety documentation.	Cosine similarity performed much better than Jaccard, detecting semantically equivalent clauses accurately.	Domain – specific study (railway rules). Results may not generalize to academic or general text. Some unmatched content reduced accuracy.
10	Sudhir Sahani (2017)	<i>Smart Cloud Document Clustering and Plagiarism Checker Using TF – IDF and Cosine Similarity</i>	Applied TF – IDF + Cosine Similarity for document duplication before cloud upload. Compared new documents with stored ones for redundancy.	Efficient for duplicate / plagiarized detection in cloud systems, reducing redundant storage.	Focused mainly on duplicate detection, not semantic plagiarism. Couldn't catch rephrased or contextually similar content. Threshold tuning impacts effectiveness.
11	Peter Coates & Frank Breitinger (2022)	<i>Identifying Document Similarity Using a Fast Estimation of</i>	Proposed a fast and approximate way to calculate Levenshtein distance using compression and signature – based methods. Used the Deflate compression algorithm and signature	Greatly reduced computation time compared to normal Levenshtein distance methods. Compression – based	Gives only an approximation, not exact results. Accuracy drops for very short or unrelated documents. Depends on the

		<i>the Levenshtein Distance Based on Compression and Signatures</i>	extraction to create compact document versions. Compared similarity by checking compressed size differences and signature overlaps.	estimates matched well with real edit distances. Works well for large – scale or near – duplicate text detection when exact distance is not required.	type and efficiency of the compression algorithm used.
12	Kensuke Baba, Tetsuya Nakatoh, Toshiro Minami (2017)	<i>Plagiarism Detection Using Document Similarity Based on Distributed Representation</i>	Proposed three plagiarism detection methods: LCS, Local LCS (LLCS), and Weighted Local LCS (WLLCS). Used word2vec embeddings to get word representations. Tested on the PAN 2013 plagiarism dataset. Evaluated with accuracy, precision, recall, and F – measure.	WLLCS gave the best results when recall $\geq 99.9\%$. Distributed word embeddings improved accuracy for complex plagiarism cases. Successfully detected even hidden or paraphrased plagiarism.	Needs parameter tuning (penalty scores for mismatch, insertion, deletion). Accuracy depends on the training corpus size and quality used in word2vec. Used only one dataset.
13	Nurhayati, Busman (2017)	<i>Development of Document Plagiarism Detection Software Using Levenshtein Distance</i>	Used Levenshtein Distance (Edit Distance) algorithm. Platform: Android (Java – based) app. Input: PDF documents (up to 40 pages). Process: compared strings using insertion, deletion, and substitution operations. Output: similarity	Successfully detected plagiarism using character – level comparison. Provided accurate similarity percentages (100% for identical docs). Efficient for comparing text – based	Works only with PDF files (no Word/txt support). Limited to 40 – page documents due to phone memory. Cannot detect semantic plagiarism (synonym use). Only shows

		<i>Algorithm on Android Smartphone</i>	percentage, distance value, number of strings, and runtime.	documents. Smaller edit distance = higher similarity.	similarity % - no clear plagiarism level.
14	Ashraf S. Hussein (2015)	<i>Arabic Document Similarity Analysis using N – Grams and Singular Value Decomposition</i>	Preprocessing: tokenization, stop – word removal, and stemming. Extracted N – grams ($n = 2 – 5$) and applied TF – IDF weighting. Used Latent Semantic Analysis (LSA) with Singular Value Decomposition (SVD).	Accurately detected both literal and semantic similarity. Performed better than existing plagiarism tools, especially for paraphrased documents. 3 – gram configuration gave the best accuracy.	Used only Arabic academic documents (30 samples). Small dataset; results may vary for larger or multilingual data. Future work: add deep learning embeddings for better semantic results.
15	M. Sangeetha , P. Keerthika, K. Devendra n, S. Sridhar, S. Shree Raagav, T. Vigneshw ar (2022)	<i>Compute Query and Document Similarity using Explicit Semantic Analysis</i>	Used Explicit Semantic Analysis (ESA), Convolutional Latent Semantic Model (CLSM), and NLP – based spell checking. Tools: Wikipedia – based ESA, Spacy.io pre-trained spell checker, and Cosine similarity. Dataset: Cranfield (1400 docs, 225 queries). Steps: preprocessing → lemmatization → spell check → ESA vectorization → cosine similarity → ranking with nDCG.	ESA – based method worked better than TF – IDF and LSA in handling synonyms and multiple meanings. Got an nDCG score of 0.32, even with limited concept data. Query autocomplete (CLSM) improved user experience. Detected deeper semantic relations.	Used only 100 Wikipedia articles for concept mapping – a larger set would improve accuracy. ESA index creation is computationally heavy. Dataset is small – needs real – world testing. Future work: expand ESA with full Wikipedia, use deep embeddings (BERT, GPT), and optimize runtime.

16	Richard Lackes, Julian Bartels, Esther Berndt, Erik Frank (2009)	<i>A Word – Frequency Based Method for Detecting Plagiarism in Documents</i>	Method: automated plagiarism detection using word frequency analysis and N – gram hashing. Steps: 1. Text preprocessing (removal of punctuation, quotes) 2. Identify rare / unique words. 3. Form search queries to find possible source docs. 4. Create N – gram fingerprints. 5. Compare hashes for plagiarism index.	Detected direct, self, and mosaic plagiarism effectively. Language – independent – doesn't require same language sources. Web integration improved detection accuracy. Using rare words reduced false matches.	Cannot detect translated or restructured plagiarism. Slow for large documents. Depends on search engine accuracy. No semantic comparison – fails to catch idea – level plagiarism. Future work: add style and semantic analysis, use ontologies, integrate modern web crawlers.
17	Ramadan Thkri Hassan, Nawzat Sadiq Ahmed (2023)	<i>Evaluating of Efficacy Semantic Similarity Methods for Comparison of Academic Thesis and Dissertation Texts</i>	Compared four similarity methods: TF – IDF, Doc2Vec, BERT, and SBERT. Used Cosine Similarity for measuring similarity. Dataset: 27 theses (Duhok Polytechnic) and 100 dissertations (ProQuest). Preprocessing: text extraction, normalization, stop – word removal, stemming, and lemmatization.	TF – IDF gave the best results in accuracy and efficiency. Outperformed advanced models like BERT and SBERT for long documents. BERT / SBERT took more time but didn't improve accuracy. For long texts, statistical methods are	Used only English documents. Small dataset (127 total). Future work: test on larger, multilingual datasets and combine TF – IDF with contextual embeddings for better scalability. Try other similarity metrics like

			Evaluated using Accuracy, Precision, Recall, F1 – score, and Time.	more reliable than contextual embeddings.	Jaccard or Manhattan distance.
18	Jon Ezeiza Alvarez (2017)	<i>A review of word embedding and document similarity algorithms applied to academic text</i>	Standard NLP preprocessing techniques are employed: tokenization, normalization (mainly Porter stemming), parsing (POS tagging), and vector space modeling, largely implemented using NLTK and Gensim libraries.	Doc2Vec is the best – performing algorithm in those tested, but only by a small margin, and is the fastest in practice. Word Movers Distance matches baselines for abstracts but is computationally infeasible for longer documents.	The absence of standardized, domain – specific human – annotated datasets for similarity benchmarks required the creation of custom evaluation sets, which may contain noise or inconsistencies.
19	Lidia Sari (2023)	<i>Cosine Similarity based Plagiarism Detection on Electronic Documents</i>	The cosine similarity metric measures similarity between vector representations of documents, where documents are vectorized using TF – IDF. The document similarity calculation involves preprocessing steps (lowercasing, tokenization), and cosine similarity computation.	Cosine similarity proves effective for detecting plagiarism, showing independence from document length and maintaining high accuracy.	The paper does not extensively discuss handling highly paraphrased or semantically altered plagiarized content beyond lexical similarity.
20	Ahmed Patel, Kav	<i>Evaluation of cheating</i>	The paper investigates various cheating and plagiarism techniques	Advanced AI – based systems incorporating	Current plagiarism systems struggle with multi – mode

	eh Bakhtiyari , Mona Taghavi (2011)	<i>detection methods in academic writings</i>	used to bypass detection in academic writings and evaluates major plagiarism detection services such as Turnitin, iThenticate, and PlagiarismDetect.	author style recognition, fuzzy logic, and semantic analysis are proposed to improve detection, especially for cross – language plagiarism and fragmented paraphrasing.	documents (mixed text, image, audio), cross – language plagiarism, and heavily paraphrased or translated texts.
--	---	---	--	---	---

4.2 Popular approaches in the existing research

- Many studies use TF – IDF with Cosine Similarity because it is fast, easy to implement, and works well for long documents.
- Deep learning models such as BERT, SBERT, SciBERT, and MT – DNN are also commonly used because they can understand the meaning of sentences and detect paraphrasing.
- Some papers use hybrid methods, where TF – IDF is used first to reduce the number of comparisons, and then BERT – based models are applied for deeper semantic checking.

4.3 Challenges found in existing research

- A major challenge is that many popular datasets contain only short sentences, but real plagiarism often occurs in long reports or thesis documents.
- Deep learning models such as BERT are accurate but require high computational power and take a long time to run.
- Traditional methods like TF – IDF work fast but cannot detect paraphrasing or meaning – level plagiarism because they depend only on matching words.
- Some papers used very small datasets or synthetic datasets, which may not truly represent real – world plagiarism cases.
- Threshold values used to decide similarity often need careful tuning, and incorrect tuning can produce false positives or false negatives.
- Most existing systems do not handle multilingual documents, code plagiarism, handwriting, images, or diagrams.

5. DATASET DESCRIPTION

For this project, we used the **MIT Plagiarism Dataset**, which is a well-known benchmark for academic plagiarism detection tasks. The dataset contains pairs of text samples labeled as similar (1) or not similar (0), covering different types of plagiarism such as exact copying, paraphrasing, and idea – level rewording.

It belongs to the academic writing domain, making it directly relevant to our project goal – detecting similarity between student reports and research documents.

- Source: MIT Plagiarism Corpus (publicly available academic dataset)
- Domain: Academic writing and research documents
- Size: Approximately 15,000 text pairs after preprocessing
- Label Meaning:
 - 1 – Text pairs that are semantically or lexically similar
 - 0 – Text pairs that are unrelated

```

A young boy wearing a bright yellow raincoat, jean shorts and sandals squats down as he looks at an older boy and girl who are standing next to him and looking down some railroad tracks. The boy is swimming 0
A young boy wearing a bright yellow raincoat, jean shorts and sandals squats down as he looks at an older boy and girl who are standing next to him and looking down some railroad tracks. The boy is wearing a yellow coat 1
a child taking a break from playing to eat an ice cream. An old dog eating a boys ice cream. 0
a child taking a break from playing to eat an ice cream. A child eating ice cream. 1
A little child eats ice cream while talking on the phone. A child eats dessert. 1
A little child eats ice cream while talking on the phone. An adult eats ice cream. 0
A young girl in a blue shirt is eating an ice cream cone. The girl is young. 1
A young girl in a blue shirt is eating an ice cream cone. The girl is wearing a red shirt 0
A young girl wearing a blue mickey mouse t-shirt eating a red, white, and blue popsicle with many different bicycles in the background. A young girl is drinking a cup of water inside. 0
A young girl wearing a blue mickey mouse t-shirt eating a red, white, and blue popsicle with many different bicycles in the background. A young girl is eating a popsicle. 1
A young child in a blue shirt holds a phone and eats something while standing in front of bicycles. An old man in a red shirt holds a phone. 0
A young child in a blue shirt holds a phone and eats something while standing in front of bicycles. A young child in a blue shirt holds a phone. 1
A mountain biker rides in the woods. A biker in the woods. 1
A mountain biker rides in the woods. A person drives a car. 0
A woman rides her bike by some trees. horse runs from bull 0
A woman rides her bike by some trees. woman on bike 1
A short, brown hair girl is assembling a pink roof cover. No small car inside pink house 0
A short, brown hair girl is assembling a pink roof cover. A small girl is fixing her pink roof cover 1
A woman assembles a white metal frame on the sidewalk. She is building a rocketship that will soon have her on mars with all the other awesome people. 0
A woman assembles a white metal frame on the sidewalk. A woman builds a something outside. 1
A young woman sands down an old shelf. A man dries his hands. 0
A young woman sands down an old shelf. A woman sands a shelf. 1
a biker bikes through the woods. A person rides a bike in the woods. 1
a biker bikes through the woods. A person runs in the woods. 0
A young woman with a pink bag and white umbrella is standing at a metro station. The woman is standing 1
A young woman with a pink bag and white umbrella is standing at a metro station. The woman is wearing no clothes 0
Man sitting with headphones at a laptop. A man with his laptop 1
Man sitting with headphones at a laptop. The man is using 25 laptops 0
A man is sitting outside at a table and his backpack is next to him on the ground. The man is standing indoors. 0

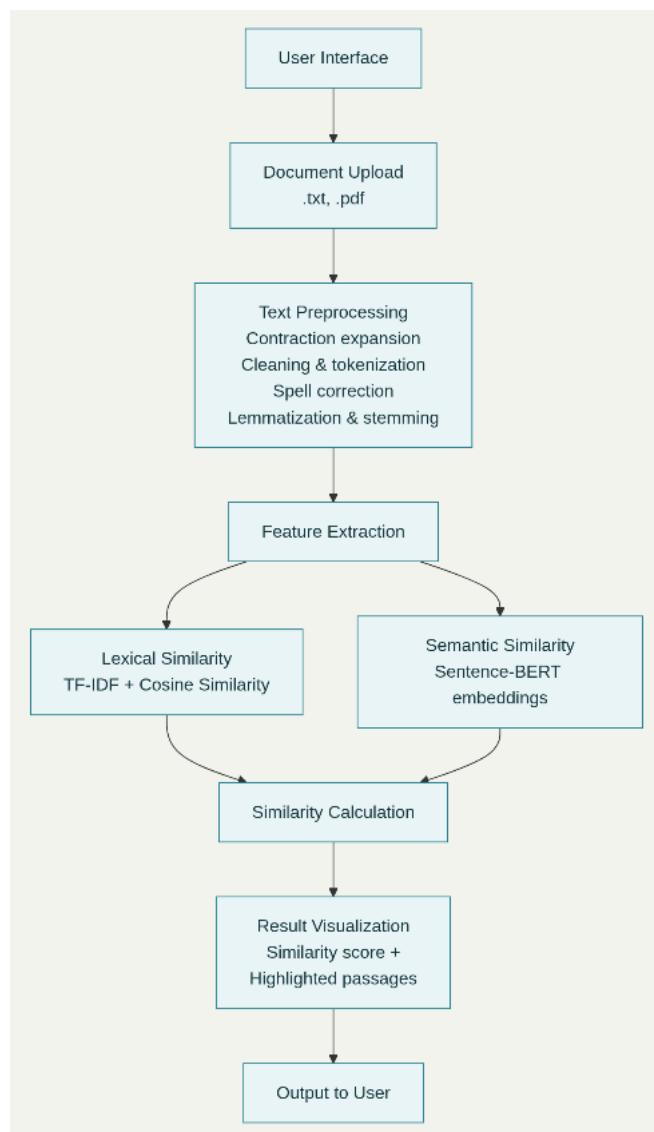
```

6. UNIQUE ASPECTS

- The project combines both fast and intelligent methods through a two – stage hybrid approach.
- TF – IDF with Cosine Similarity is used first to quickly filter and detect directly copied or highly similar documents.
- Then, deep learning models – BERT – base – NLI – mean – tokens and SBERT variants (all – MiniLM - L6 - v2 and all – mpnet – base - v2) – analyse the filtered results to identify paraphrased and meaning-level similarities that traditional methods might miss.

- The models were fine – tuned on the MIT Plagiarism Dataset, allowing them to learn real academic writing styles instead of synthetic general purpose text.
- For longer reports, the system can process the text in segments or chunks, improving accuracy in comparing large academic documents.
- The inclusion of lightweight SBERT models helps reduce computation time while maintaining high accuracy – making the system more practical for regular use.
- The design allows for clear output, showing both word – level and semantic similarity scores, which helps users and teachers understand how the documents are related.

7. ARCHITECTURE DIAGRAM



8. PROPOSED MODEL / APPROACH

The proposed system aims to detect both direct and paraphrased plagiarism in academic writing by combining traditional NLP techniques with advanced deep learning models. The approach that balances speed, accuracy, and semantic understanding.

Stage 1: Text Preprocessing

Before comparison, the input documents are cleaned and normalized using the following steps:

- Lowercasing, removing punctuation, URLs, and extra spaces
- Tokenization, stopword removal, lemmatization, and stemming
- Contraction expansion

These steps help ensure that the models focus only on meaningful words and reduce noise in the text.

Stage 2: Traditional Similarity Detection (TF – IDF + Cosine Similarity)

In this stage, both documents are represented as TF – IDF vectors, which capture how important each word is within a document.

Then, Cosine Similarity measures the angle between these vectors – if the angle is small (value close to 1), the documents are similar.

This step helps to quickly detect directly copied or lexically similar content, filtering out unrelated pairs early.

Stage 3: Deep Semantic Similarity (BERT + SBERT Models)

For deeper understanding beyond surface – level word matches, the system uses transformer – based models:

- BERT – base – NLI – mean – tokens
- SBERT all – MiniLM - L6 -v2
- SBERT all – mpnet – base - v2

These models convert sentences into semantic embeddings, capturing meaning and context. By computing cosine similarity between these embeddings, the system can detect rephrased or conceptually similar content that traditional methods often miss. Each model was fine – tuned to adapt it specifically for academic text.

Stage 4: Combined Similarity Score and Decision

The final similarity score is a weighted combination of:

- The TF – IDF cosine score (for direct overlap)
- The BERT / SBERT cosine score (for semantic similarity)

Based on a similarity threshold (e.g., 0.5), the system classifies the document pair as plagiarized (1) or non – plagiarized (0).

8.2 Advantages of the Proposed Approach

- Hybrid design: Combines speed (TF – IDF) and intelligence (BERT / SBERT)
- Domain – specific learning: Fine – tuned on academic data
- Handles paraphrasing: Detects plagiarism even when wording changes
- Scalable: Works for full documents, not just short sentences
- Interpretable: Produces both direct and semantic similarity scores

8.3 How the project addresses research gaps

Most existing plagiarism detection systems focus only on one type of method - either traditional or deep learning — which limits their performance. Traditional ones are fast but only catch exact word overlaps and are limited as they are not considering the meaning behind the context, while deep models understand meaning but are slow and often not trained for academic writing. Our project fills this gap by combining both traditional and advanced models for a complete and practical solution.

8.4 Relevance and Challenges of the Proposed Approach

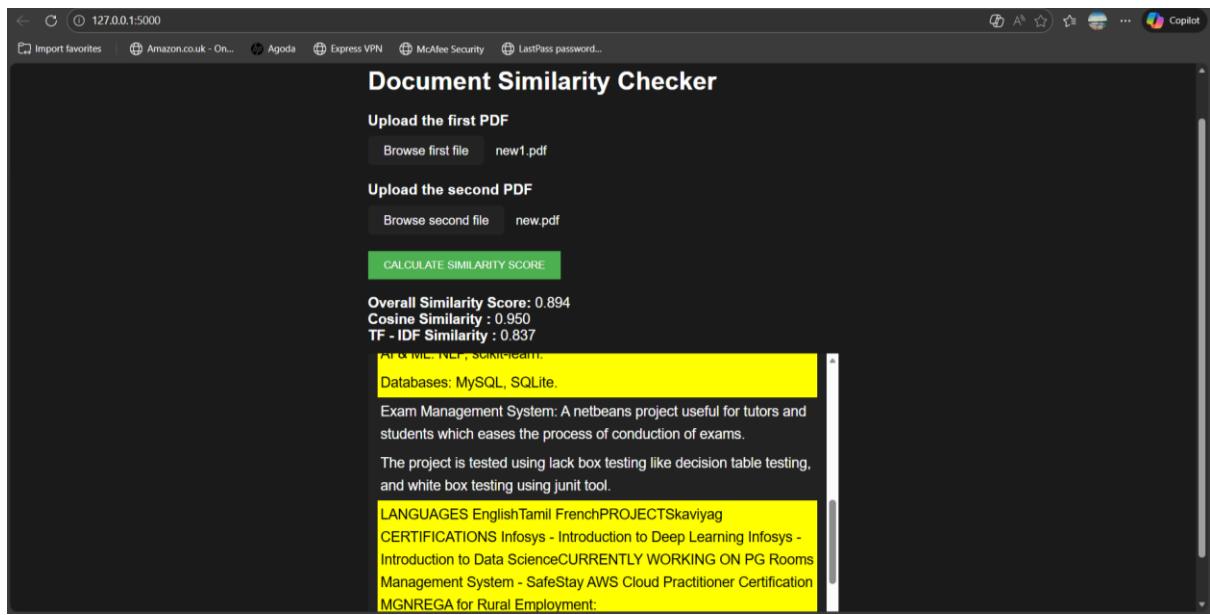
The MIT Plagiarism Dataset is highly relevant because it mirrors the types of text reuse found in real academic submissions – making it ideal for building and evaluating a Document Similarity Checker.

A key challenge was handling text imbalance (more non – similar than similar pairs) and processing large text lengths efficiently. These were managed by careful data balancing, text chunking, and the use of lightweight transformer models to optimize speed and memory usage.

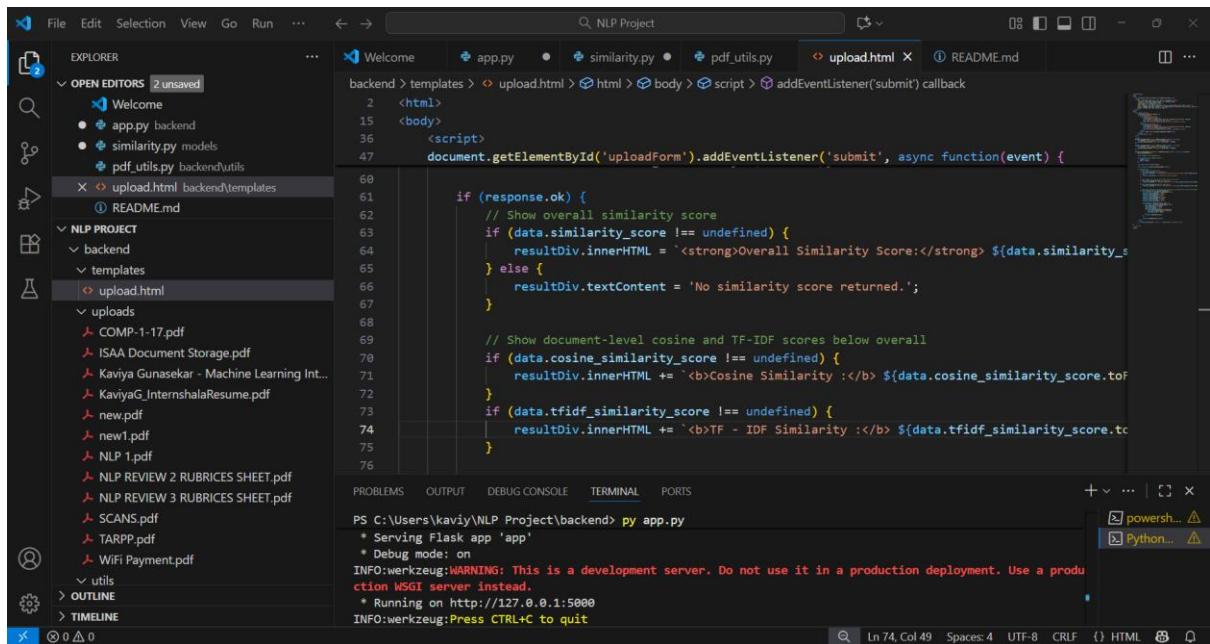
9. IMPLEMENTATION OF THE APPROACH / MODELS

9.1 BASELINE MODEL

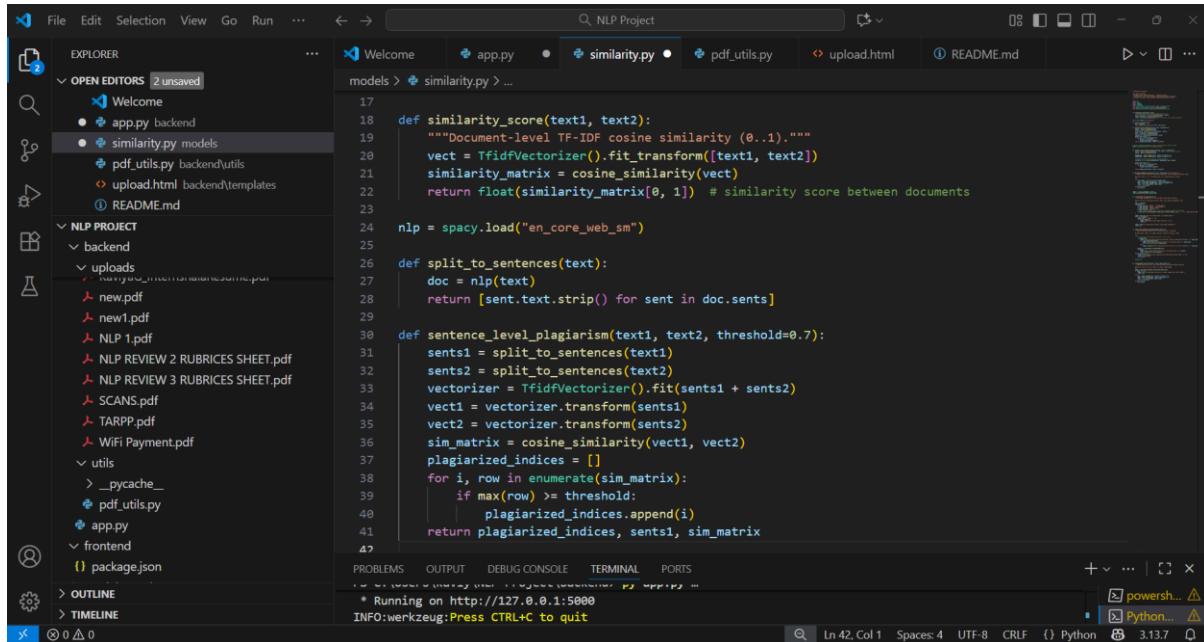
The baseline model consists of three main components: a Flask web application, a backend similarity engine, and a frontend user interface. The Flask application exposes a simple UI and an /upload endpoint that accepts two PDF files, processes them, and returns a JSON response with similarity scores and sentence – level highlights. The backend uses utilities for document and sentence – level similarity, combining TF – IDF to compute both document – level and per – sentence similarity scores. The frontend provides an intuitive upload form, displays the overall similarity score, and visually highlights plagiarized sentences in the extracted text, making it easy for users to interpret the results. This architecture ensures a seamless workflow from file upload to similarity analysis and result visualization, supporting academic writing and plagiarism detection tasks efficiently.



The uploaded pdf files are saved locally.



Splitting the document into individual sentences and calculating their similarity.

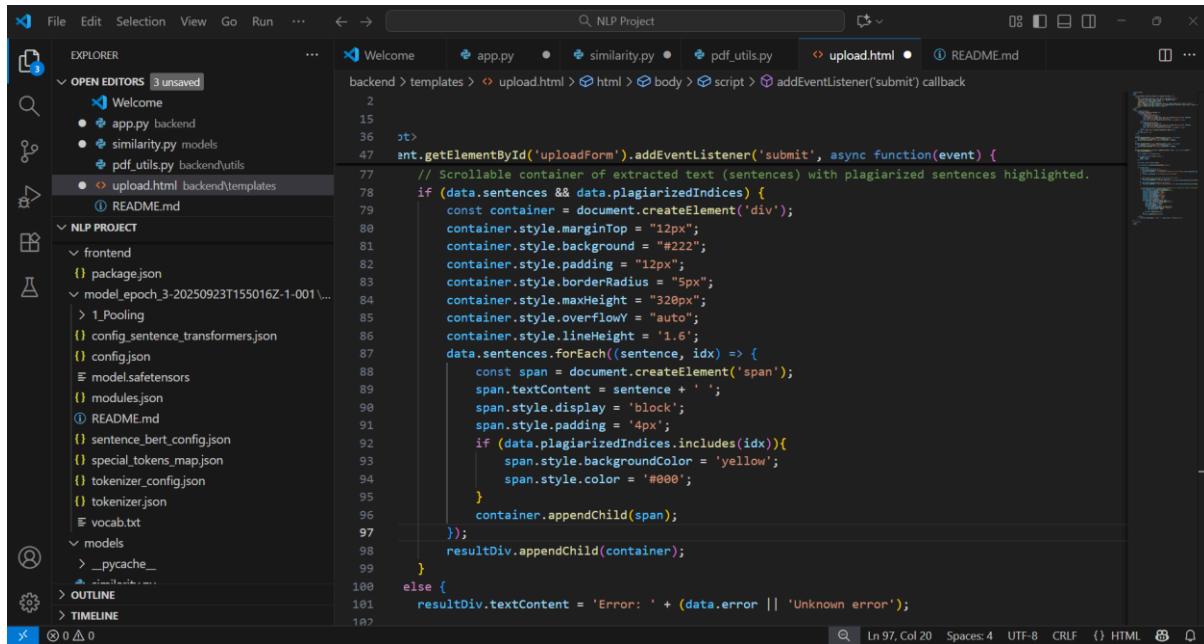


The screenshot shows a code editor interface with the title "NLP Project". The left sidebar displays the project structure under "NLP PROJECT" and "OPEN EDITORS". The main editor area contains Python code for calculating sentence similarity:

```
17 def similarity_score(text1, text2):
18     """Document-level TF-IDF cosine similarity (0..1)."""
19     vect = TfidfVectorizer().fit_transform([text1, text2])
20     similarity_matrix = cosine_similarity(vect)
21     return float(similarity_matrix[0, 1]) # similarity score between documents
22
23 nlp = spacy.load("en_core_web_sm")
24
25 def split_to_sentences(text):
26     doc = nlp(text)
27     return [sent.text.strip() for sent in doc.sents]
28
29 def sentence_level_plagiarism(text1, text2, threshold=0.7):
30     sents1 = split_to_sentences(text1)
31     sents2 = split_to_sentences(text2)
32     vectorizer = TfidfVectorizer().fit(sents1 + sents2)
33     vect1 = vectorizer.transform(sents1)
34     vect2 = vectorizer.transform(sents2)
35     sim_matrix = cosine_similarity(vect1, vect2)
36     plagiarized_indices = []
37     for i, row in enumerate(sim_matrix):
38         if max(row) >= threshold:
39             plagiarized_indices.append(i)
40     return plagiarized_indices, sents1, sim_matrix
41
42
```

The status bar at the bottom indicates "INFO:werkzeug:Press CTRL+C to quit".

Displaying a scrollable field of extracted text where the similar sentences are highlighted.



The screenshot shows a code editor interface with the title "NLP Project". The left sidebar displays the project structure under "NLP PROJECT" and "OPEN EDITORS". The main editor area contains JavaScript code for a file upload form:

```
2
15
36
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
```

The status bar at the bottom indicates "INFO:werkzeug:Press CTRL+C to quit".

9.2 BERT MODEL

SaveCheckpointCallback is used to save the model at the end of each training epoch.

```
[ ]    from sentence_transformers import SentenceTransformer, InputExample, losses, evaluation
from torch.utils.data import DataLoader
from sentence_transformers import models, LoggingHandler
import os

class SaveCheckpointCallback:
    def __init__(self, model, save_path):
        self.model = model
        self.save_path = save_path

    def on_epoch_end(self, epoch):
        # Save the model checkpoint at the end of each epoch
        path = self.save_path.format(epoch=epoch)
        self.model.save(path)
        print(f"Model saved at {path}")

    # Load the pre-trained bert-base-nli-mean-tokens model
model = SentenceTransformer('bert-base-nli-mean-tokens')

# Prepare your training examples assuming train_data is a list of tuples (doc1, doc2, label)
train_examples = [InputExample(texts=[d[0], d[1]], label=float(d[2])) for d in train_data]

# Create a DataLoader to batch your training samples
train_dataloader = DataLoader(train_examples, shuffle=True, batch_size=16)

# Define a loss function suitable for sentence embeddings (e.g., cosine similarity loss)
train_loss = losses.CosineSimilarityLoss(model)

checkpoint_callback = SaveCheckpointCallback(model, '/content/drive/MyDrive/NLP/model_epoch_{epoch}' )
```

```
[ ]    checkpoint_callback = SaveCheckpointCallback(model, '/content/drive/MyDrive/NLP/model_epoch_{epoch}')

# Start training with callback for checkpoints
num_epochs = 3

# Fine-tune the model with your data; define number of epochs
for epoch in range(num_epochs):
    model.fit(train_objectives=[(train_dataloader, train_loss)], epochs=1, warmup_steps=100)
    checkpoint_callback.on_epoch_end(epoch + 1)

# After training, you can encode documents as usual
doc1_embs = model.encode([d[0] for d in train_data], convert_to_tensor=True)
doc2_embs = model.encode([d[1] for d in train_data], convert_to_tensor=True)
```

▼

[751/751 01:48, Epoch 1/1]

Step	Training Loss
500	0.154000

Model saved at /content/drive/MyDrive/NLP/model_epoch_1

[751/751 01:47, Epoch 1/1]

Step	Training Loss
500	0.080200

Model saved at /content/drive/MyDrive/NLP/model_epoch_2

[751/751 01:47, Epoch 1/1]

```

Step Training Loss
500 0.034600
Model saved at /content/drive/MyDrive/NLP/model_epoch_3

[ ] from sklearn.metrics import accuracy_score, precision_recall_fscore_support
import torch

# Predict similarity on validation/test data
model.eval()

y_true = [float(d[2]) for d in val_data]
y_pred = []

for doc1, doc2, _ in val_data:
    emb1 = model.encode(doc1, convert_to_tensor=True)
    emb2 = model.encode(doc2, convert_to_tensor=True)
    sim = torch.nn.functional.cosine_similarity(emb1, emb2, dim=0).item()
    y_pred.append(1 if sim > 0.5 else 0)

precision, recall, f1, _ = precision_recall_fscore_support(y_true, y_pred, average='binary')
accuracy = accuracy_score(y_true, y_pred)

print(f"Accuracy: {accuracy}, Precision: {precision}, Recall: {recall}, F1 Score: {f1}")

Accuracy: 0.7802929427430093, Precision: 0.7769230769230769, Recall: 0.7952755905511811, F1 Score: 0.7859922178988327

```

Sets up callback to save model checkpoints in Google Drive with epoch number in path.

My Drive > NLP ▾

Type ▾ People ▾ Modified ▾ Source ▾

Name	Owner	Date modified	File size	⋮
model_epoch_1	(K) me	Sep 23 me	—	⋮
model_epoch_2	(K) me	Sep 23 me	—	⋮
model_epoch_3	(K) me	Nov 6 me	—	⋮
3.txt	(K) me	Nov 6 me	236 bytes	⋮
4.txt	(K) me	Nov 6 me	195 bytes	⋮
new.txt	(K) me	Nov 6 me	2 KB	⋮
new1.txt	(K) me	Nov 6 me	2 KB	⋮
one.txt	(K) me	Nov 6 me	243 bytes	⋮
original.txt	(K) me	Sep 23 me	1.5 MB	⋮
preprocessed_dataset.txt	(K) me	Sep 23 me	966 KB	⋮

Loading the final saved epoch.

```
[ ] ➜ !pip install contractions
from google.colab import drive
drive.mount('/content/drive')

from sentence_transformers import SentenceTransformer
import torch
import re, nltk, contractions
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer, PorterStemmer
from nltk.tokenize import word_tokenize, sent_tokenize

# --- load trained model ---
model = SentenceTransformer('/content/drive/MyDrive/NLP/model_epoch_3')

...
*** Collecting contractions
  Downloading contractions-0.1.73-py2.py3-none-any.whl.metadata (1.2 kB)
Collecting textsearch>=0.0.21 (from contractions)
  Downloading textsearch-0.0.24-py2.py3-none-any.whl.metadata (1.2 kB)
Collecting anyascii (from textsearch>0.0.21->contractions)
  Downloading anyascii-0.3.3-py3-none-any.whl.metadata (1.6 kB)
Collecting pyahocorasick (from textsearch>0.0.21->contractions)
  Downloading pyahocorasick-2.2.0-cp312-cp312-manylinux_2_17_x86_64_manylinux2014_x86_64.whl.metadata (13 kB)
Downloading contractions-0.1.73-py2.py3-none-any.whl (8.7 kB)
Downloading textsearch-0.0.24-py2.py3-none-any.whl (7.6 kB)
Downloading anyascii-0.3.3-py3-none-any.whl (345 kB)
  345.1/345.1 kB 11.7 MB/s eta 0:00:00
Downloading pyahocorasick-2.2.0-cp312-cp312-manylinux_2_17_x86_64_manylinux2014_x86_64.whl (114 kB)
  114.9/114.9 kB 4.3 MB/s eta 0:00:00
Installing collected packages: pyahocorasick, anyascii, textsearch, contractions
Successfully installed anyascii-0.3.3 contractions-0.1.73 pyahocorasick-2.2.0 textsearch-0.0.24
Mounted at /content/drive
```

Output

The combined score is a value of 40% of the TF – IDF similarity score and 60% of the BERT similarity score.

```
[ ] ➜ doc1_clean = preprocess_text(doc1)
doc2_clean = preprocess_text(doc2)

# --- encode & compute similarity ---
emb1 = model.encode(doc1_clean, convert_to_tensor=True)
emb2 = model.encode(doc2_clean, convert_to_tensor=True)

similarity = torch.nn.functional.cosine_similarity(emb1, emb2, dim=0).item()

vectorizer = TfIdfVectorizer()
tfidf_matrix = vectorizer.fit_transform([doc1_clean, doc2_clean])
tfidf_sim = cosine_similarity(tfidf_matrix[0:1], tfidf_matrix[1:2])[0][0]

# --- ❷ BERT Similarity ---
emb1 = model.encode(doc1_clean, convert_to_tensor=True)
emb2 = model.encode(doc2_clean, convert_to_tensor=True)
bert_sim = torch.nn.functional.cosine_similarity(emb1, emb2, dim=0).item()

# --- ❸ Combine (Weighted Average) ---
# You can adjust weights depending on which you trust more
combined_sim = (0.4 * tfidf_sim) + (0.6 * bert_sim)

# --- Output ---
print(f"TF-IDF Similarity: {tfidf_sim:.4f}")
print(f"BERT Similarity: {bert_sim:.4f}")
print(f"Combined Score: {combined_sim:.4f}")

...
*** TF-IDF Similarity: 0.0497
BERT Similarity: 0.5559
Combined Score: 0.3534
```

9.3 SBERT MODELS FROM HUGGING FACE

```
import pandas as pd # Import pandas for data manipulation and analysis
import torch # Import PyTorch for tensor operations and deep learning
import re, nltk, contractions # Import regex for text pattern matching, nltk for natural language processing, and contractions for expanding
from nltk.corpus import stopwords # Import stopwords from nltk for removing common words
from nltk.stem import WordNetLemmatizer, PorterStemmer # Import lemmatizer and stemmer from nltk for text normalization
from nltk.tokenize import word_tokenize # Import word_tokenize from nltk for splitting text into words
from sklearn.model_selection import train_test_split # Import train_test_split for dividing data into training and testing sets
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score # Import various metrics for model evaluation
from sentence_transformers import SentenceTransformer, InputExample, losses # Import SentenceTransformer for creating sentence embeddings, I
from torch.utils.data import DataLoader # Import DataLoader for efficient batching of data
import matplotlib.pyplot as plt # Import matplotlib for plotting and visualization
```

1. Mount Google Drive:

```
from google.colab import drive
drive.mount('/content/drive')
```

This is necessary to access your dataset and save fine-tuned models.

2. Install Required Python Packages:

Run the following command to install `contractions` (for text normalization), `python-docx` (for DOCX file processing), and `PyMuPDF` (for PDF file processing):

```
!pip install contractions python-docx PyMuPDF
```

PREPROCESSING STEPS

- Contraction Expansion:** Converts short forms (“*isn’t*” → “*is not*”) to maintain consistency.
- Noise Removal:** Removes URLs, emails, and web – specific elements using regex.
- Whitespace Normalization:** Replaces multiple spaces and tabs with a single space.
- Lowercasing:** Converts all text to lowercase to avoid case sensitivity.
- Special Character Removal:** Removes punctuation and symbols, keeping only letters, digits, and spaces.
- Tokenization:** Splits text into sentences and individual words for analysis.
- Spell Correction:** Fixes misspelled words using a spellchecker module.
- Stop Word Removal:** Removes common words like “the”, “is”, and “and” to reduce noise.
- Lemmatization:** Converts words to their base dictionary form (“*running*” → “*run*”).
- Stemming:** Reduces words to their root form (“*connection*” → “*connect*”).



```

i CODE i TEST
stop_words = set(stopwords.words('english')) # Define a set of English stop words for efficient lookup
lemmatizer = WordNetLemmatizer() # Initialize the WordNet Lemmatizer
stemmer = PorterStemmer() # Initialize the Porter Stemmer

def preprocess_text(text):
    text = contractions.fix(str(text)) # Expand contractions (e.g., "don't" to "do not")
    text = re.sub(r"HTTP\S+|WWW\S+|HTTPS\S+", '', text) # Remove URLs
    text = re.sub(r'\S+@\S+', '', text) # Remove email addresses
    text = re.sub(r'\s+', ' ', text).strip() # Normalize whitespace by replacing multiple spaces with a single space and stripping leading/trailing spaces
    text = text.lower() # Convert text to lowercase
    text = re.sub(r'[^a-zA-Z\s]', ' ', text) # Remove special characters, keeping only alphanumeric and spaces
    words = word_tokenize(text) # Tokenize the text into individual words
    # Lemmatize, stem, and remove stop words from the tokens
    processed_words = [stemmer.stem(lemmatizer.lemmatize(w)) for w in words if w not in stop_words]
    return ' '.join(processed_words) # Join the processed words back into a single string

```

Loading Dataset.

```
# =====#
# 3) Load Dataset
# =====#
file_path = '/content/drive/MyDrive/Colab Notebooks/dataset.txt' # Define the path to the dataset file

df = pd.read_csv(file_path, sep='\t', header=None, names=['text1', 'text2', 'label']) # Load the dataset into a pandas DataFrame, specifying tab as the separator, no header, and custom column names
print(f"Loaded {len(df)} examples") # Print a confirmation message with the number of loaded examples
print(df.head()) # Display the first few rows of the DataFrame

# Optional preprocessing (if not already cleaned)
df['text1'] = df['text1'].apply(preprocess_text) # Apply the preprocess_text function to 'text1' column for cleaning and normalization
df['text2'] = df['text2'].apply(preprocess_text) # Apply the preprocess_text function to 'text2' column for cleaning and normalization
```

Loaded 15021 examples

	text1	text2	label
0	A young boy wearing a bright yellow raincoat, ...	The boy is swimming	0
1	A young boy wearing a bright yellow raincoat, ...	The boy is wearing a yellow coat	1
2	a child taking a break from playing to eat an ...	An old dog eating a boys ice cream.	0
3	a child taking a break from playing to eat an ...	A child eating ice cream.	1
4	a little child eats ice cream while talking on...	A child eats dessert.	1

Train Test Splitting.

```
# =====#
# 4 Train / Validation / Test Split
# =====#
# Split the dataset into training and a temporary set (validation + test)
train_df, temp_df = train_test_split(df, test_size=0.2, random_state=42) # test_size=0.2 means 20% for temp_df, random_state for reproducibility
# Split the temporary set into validation and test sets
val_df, test_df = train_test_split(temp_df, test_size=0.5, random_state=42) # test_size=0.5 means 50% of temp_df for test_df, 50% for val_df

print(f"Train: {len(train_df)} | Val: {len(val_df)} | Test: {len(test_df)}") # Print the number of examples in each split
```

• Train: 12016 | Val: 1502 | Test: 1503

Model Comparison Setup.

```
# =====#
# 5 Model Comparison Setup
# =====#
# Define a list of pre-trained Sentence-BERT models to be tested and fine-tuned
models_to_test = [
    'all-MiniLM-L6-v2',
    'paraphrase-MiniLM-L12-v2',
    'all-mpnet-base-v2'
]

results = {} # Initialize an empty dictionary to store evaluation results for each model
```

Convert Data for SBERT Training.

```
# =====#
# 6 Convert Data for Sentence-BERT Training
# =====#
# Create InputExample objects for the training set. Each example consists of two texts and a label (similarity score).
train_samples = [InputExample(texts=[r.text1, r.text2], label=float(r.label)) for r in train_df.iterrows()]
# Create validation pairs as a list of tuples (text1, text2, label) for evaluation.
val_pairs = [(r.text1, r.text2, r.label) for r in val_df.iterrows()]
```

```

    import os
    os.environ["WANDB_DISABLED"] = "true" # Disable Weights & Biases logging to prevent automatic logging if W&B is installed

    # Redefine models to test (if this was done in a previous cell, it's good to keep it consistent or updated here)
    models_to_test = [
        'all-MiniLM-L6-v2',      # A fast, lightweight, and accurate pre-trained Sentence-BERT model
        'all-mpnet-base-v2'      # A more powerful pre-trained Sentence-BERT model known for strong semantic understanding
    ]

    EPOCHS = 2    # Number of training epochs for fine-tuning, chosen for a good tradeoff with the dataset size
    BATCH_SIZE = 16 # Batch size for the DataLoader during training

    for model_name in models_to_test:
        print(f"\n➡️ Training Model: {model_name}") # Inform the user which model is currently being trained
        model = SentenceTransformer(model_name) # Initialize the SentenceTransformer model with the pre-trained weights

        # Create DataLoader and define loss function for training
        train_dataloader = DataLoader(train_samples, shuffle=True, batch_size=BATCH_SIZE) # DataLoader to efficiently feed training samples in batches
        train_loss = losses.CosineSimilarityLoss(model) # Use CosineSimilarityLoss to maximize similarity for positive pairs and minimize for negative pairs

        # Fine-tune the model
        model.fit(
            train_objectives=[(train_dataloader, train_loss)], # Specify the DataLoader and loss function for training
            epochs=EPOCHS, # Set the number of training epochs
            warmup_steps=100, # Number of steps for linear warmup of the learning rate
            show_progress_bar=True # Display a progress bar during training
        )

        # Save the fine-tuned model to Google Drive
        save_path = f"/content/drive/MyDrive/NLP/{model_name.replace('/', '_')}_fine_tuned" # Define the path to save the model, replacing '/' for valid directory names
        model.save(save_path) # Save the model components (config, tokenizer, weights)
        print(f"✅ Model saved at {save_path}") # Confirm model saving location

    # --- Evaluation on Validation Set ---
    y_true, y_pred = [], [] # Initialize lists to store true labels and predicted labels
    for text1, text2, label in val_pairs:
        emb1 = model.encode(text1, convert_to_tensor=True) # Encode the first text into a tensor embedding
        emb2 = model.encode(text2, convert_to_tensor=True) # Encode the second text into a tensor embedding
        sim = torch.nn.functional.cosine_similarity(emb1, emb2, dim=0).item() # Calculate cosine similarity between the two embeddings
        pred = 1 if sim > 0.5 else 0 # Predict 1 (similar) if similarity is above 0.5, else 0 (dissimilar)
        y_true.append(label) # Add the true label to the list
        y_pred.append(pred) # Add the predicted label to the list

    # Compute evaluation metrics
    acc = accuracy_score(y_true, y_pred) # Calculate overall accuracy
    f1 = f1_score(y_true, y_pred) # Calculate F1-score, a harmonic mean of precision and recall
    prec = precision_score(y_true, y_pred) # Calculate precision: true positives / (true positives + false positives)
    rec = recall_score(y_true, y_pred) # Calculate recall: true positives / (true positives + false negatives)

    results[model_name] = {'Accuracy': acc, 'F1': f1, 'Precision': prec, 'Recall': rec} # Store metrics in the results dictionary
    print(f"📊 Accuracy: {acc:.4f} | F1: {f1:.4f} | Precision: {prec:.4f} | Recall: {rec:.4f}") # Print the computed metrics for the current model

```

- all – MiniLM - L6 - v2: A lightweight and efficient model.
- all – mpnet – base - v2: A more powerful model known for strong semantic understanding.

Training Process

- Loss Function:** CosineSimilarityLoss is used, which aims to maximize the cosine similarity between the embeddings of similar sentence pairs (label 1) and minimize it for dissimilar pairs (label 0).
- Epochs:** Each model is fine – tuned for 2 epochs.
- Batch Size:** A batch size of 16 is used during training.
- Warmup Steps:** A warmup period of 100 steps helps stabilize training at the beginning.

Evaluation

After fine – tuning, each model is evaluated on a dedicated validation set (val_df). The evaluation process involves encoding text pairs from the validation set, calculating the cosine similarity between their embeddings, and classifying them as similar (if similarity > 0.5) or dissimilar. The following metrics are then computed to assess model performance:

- **Accuracy:** The proportion of correctly classified pairs.
- **F1 – score:** The harmonic mean of precision and recall, useful for imbalanced datasets.
- **Precision:** The proportion of true positive results among all positive results (true and false positives).
- **Recall:** The proportion of true positive results among all actual positive results (true positives and false negatives).

all – MiniLM - L6 - v2 _ fine _ tuned model and results.

```
results[model_name] = {'Accuracy': acc, 'F1': f1, 'Precision': prec, 'Recall': rec} # Store metrics in the results dictionary
print(f"\n Accuracy: {acc:.4f} | F1: {f1:.4f} | Precision: {prec:.4f} | Recall: {rec:.4f}") # Print the computed metrics for the current model

# Training Model: all-MiniLM-L6-v2
Using the 'WANDB_DISABLE' environment variable is deprecated and will be removed in v5. Use the --report_to flag to control the integrations used for logging result (for instance --report_to wandb)
Using the 'WANDB_DISABLE' environment variable is deprecated and will be removed in v5. Use the --report_to flag to control the integrations used for logging result (for instance --report_to wandb)

[1502/1502 01:06, Epoch 2/2]
Step Training Loss
500 0.165200
1000 0.140700
1500 0.134400
Model saved at /content/drive/MyDrive/NLP/all-MiniLM-L6-v2_fine_tuned
Accuracy: 0.8222 | F1: 0.8320 | Precision: 0.7993 | Recall: 0.8675
```

all – mpnet – base - v2 _ fine _ tuned model and results.

```
# Training Model: all-mpnet-base-v2
Using the 'WANDB_DISABLE' environment variable is deprecated and will be removed in v5. Use the --report_to flag to control the integrations used for logging result (for instance --report_to wandb)
Using the 'WANDB_DISABLE' environment variable is deprecated and will be removed in v5. Use the --report_to flag to control the integrations used for logging result (for instance --report_to wandb)

[1222/1502 03:08 <0:04:43, 6.47 it/s, Epoch 1.63/2]
Step Training Loss
500 0.149700
1000 0.121800
[1502/1502 03:50, Epoch 2/2]
Step Training Loss
500 0.149700
1000 0.121800
1500 0.108800
Model saved at /content/drive/MyDrive/NLP/all-mpnet-base-v2_fine_tuned
Accuracy: 0.8296 | F1: 0.8369 | Precision: 0.8131 | Recall: 0.8622
```

Comparing the models and testing similarity.

```
# COMPARING THE MODELS AND TESTING THE SIMILARITY BETWEEN THE DOCUMENTS
from sentence_transformers import SentenceTransformer
import torch

# --- Load both models (paths to your saved fine-tuned versions) ---
# Initialize a dictionary to store the loaded fine-tuned models for easy access
models = {
    # Load the fine-tuned 'all-MiniLM-L6-v2' model from the specified Google Drive path
    "all-MiniLM-L6-v2": SentenceTransformer('/content/drive/MyDrive/NLP/all-MiniLM-L6-v2_fine_tuned'),
    # Load the fine-tuned 'all-mpnet-base-v2' model from the specified Google Drive path
    "all-mpnet-base-v2": SentenceTransformer('/content/drive/MyDrive/NLP/all-mpnet-base-v2_fine_tuned')
}

# =====
# Upload 2 Documents (TXT, DOCX, or PDF)
# Compare with Both Fine-Tuned Models
# =====

from google.colab import files
from sentence_transformers import SentenceTransformer
import torch
import os

# Install text extractors (run once): 'python-docx' for .docx files and 'PyMuPDF' for .pdf files
!pip install python-docx PyMuPDF > /dev/null

from docx import Document # Import Document class for reading .docx files
import fitz # PyMuPDF is imported as fitz for reading .pdf files

# --- Helper functions to read different file types ---
def read_txt(path):
    # Function to read text from a .txt file
    with open(path, 'r', encoding='utf-8', errors='ignore') as f:
        return f.read()

def read_docx(path):
    # Function to read text from a .docx file by iterating through paragraphs
    doc = Document(path)
    return '\n'.join([p.text for p in doc.paragraphs])
```

```

def read_pdf(path):
    # Function to read text from a .pdf file page by page using PyMuPDF
    text = ""
    with fitz.open(path) as pdf:
        for page in pdf:
            text += page.get_text()
    return text

def read_file_auto(path):
    # Automatically detects file type based on extension and calls the appropriate reader function
    ext = os.path.splitext(path)[1].lower()
    if ext == '.txt':
        return read_txt(path)
    elif ext == '.docx':
        return read_docx(path)
    elif ext == '.pdf':
        return read_pdf(path)
    else:
        raise ValueError(f"Unsupported file type: {ext}")

# --- Upload your documents ---
print(" Please upload any two files (.txt, .docx, or .pdf)")
uploaded = files.upload() # Prompt the user to upload files

file_names = list(uploaded.keys())
if len(file_names) < 2:
    raise ValueError("Please upload exactly two documents.")

# --- Upload your documents ---|
print(" Please upload any two files (.txt, .docx, or .pdf)")
uploaded = files.upload() # Prompt the user to upload files

file_names = list(uploaded.keys())
if len(file_names) < 2:
    raise ValueError("Please upload exactly two documents.")

# --- Read both files (auto-detect type) ---
doc1 = read_file_auto(file_names[0]) # Read the content of the first uploaded file
doc2 = read_file_auto(file_names[1]) # Read the content of the second uploaded file
print(f"\n Uploaded and extracted text from: {file_names[0]}, {file_names[1]}")

# --- Load fine-tuned models from Drive ---
# Re-loading the fine-tuned Sentence-BERT models from Google Drive for comparison
models = {
    "all-MiniLM-L6-v2": SentenceTransformer('/content/drive/MyDrive/NLP/all-MiniLM-L6-v2_fine_tuned'),
    "all-mnpt-base-v2": SentenceTransformer('/content/drive/MyDrive/NLP/all-mnpt-base-v2_fine_tuned')
}
print(" Fine-tuned models loaded successfully.\n")

# --- Preprocess both docs (reuse your existing preprocess_text function) ---
doc1_clean = preprocess_text(doc1) # Apply the preprocessing function to the first document
doc2_clean = preprocess_text(doc2) # Apply the preprocessing function to the second document

# --- Compare similarity for both models ---
for name, model in models.items(): # Iterate through each loaded model
    emb1 = model.encode(doc1_clean, convert_to_tensor=True) # Encode the first cleaned document into an embedding tensor
    emb2 = model.encode(doc2_clean, convert_to_tensor=True) # Encode the second cleaned document into an embedding tensor

```

```

# --- Compare similarity for both models ---
for name, model in models.items(): # Iterate through each loaded model
    emb1 = model.encode(doc1_clean, convert_to_tensor=True) # Encode the first cleaned document into an embedding tensor
    emb2 = model.encode(doc2_clean, convert_to_tensor=True) # Encode the second cleaned document into an embedding tensor
    sim = torch.nn.functional.cosine_similarity(emb1, emb2, dim=0).item() # Calculate the cosine similarity between the two embeddings

    print(f"Model: {name}") # Print the name of the current model
    print(f"Cosine Similarity: {sim:.4f}") # Print the calculated cosine similarity

    # Interpret the similarity score for plagiarism detection
    if sim > 0.75:
        print("Likely plagiarized or highly similar.\n") # High similarity indicates likely plagiarism
    elif sim > 0.5:
        print("Moderately similar (possible paraphrasing).\n") # Moderate similarity suggests paraphrasing or related topics
    else:
        print("Not similar.\n") # Low similarity indicates distinct content

```

Output

```

... Please upload any two files (.txt, .docx, or .pdf)
Choose Files 2 files
docu1.txt.docx(application/vnd.openxmlformats-officedocument.wordprocessingml.document) - 17483 bytes, last modified: 8/5/2025 - 100% done
docu2.txt.docx(application/vnd.openxmlformats-officedocument.wordprocessingml.document) - 15117 bytes, last modified: 8/5/2025 - 100% done
Saving docu1.txt.docx to docu1.txt.docx
Saving docu2.txt.docx to docu2.txt.docx

Uploaded and extracted text from: docu1.txt.docx, docu2.txt.docx
Fine-tuned models loaded successfully.

• Model: all-MiniLM-L6-v2
Cosine Similarity: 0.3595
Not similar.

• Model: all-mpnet-base-v2
Cosine Similarity: 0.1842
Not similar.

```

Working of the Selected Models

1. TF – IDF with Cosine Similarity (Baseline Model)

- Term Frequency – Inverse Document Frequency converts text into numerical vectors.
 - *TF* measures how often a word appears in a document.
 - *IDF* reduces the weight of common words that appear in many documents.
- Each document becomes a vector of word importance scores.
- Cosine Similarity then measures the angle between two document vectors
 - If the angle is small (cosine value close to 1), the documents are similar.
 - This works well for detecting exact or near – exact word matches but not deep meaning.

2. BERT – base – NLI – mean – tokens

- BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model based on the Transformer architecture.
- It reads text in both directions (left – to – right and right – to – left) to understand context – unlike older models that read only one way.
- Each word in a sentence is represented as a contextual embedding – its meaning depends on the surrounding words.
 - Example: The word “*bank*” means different things in “*river bank*” and “*money bank*”.
- The “NLI – mean – tokens” version is pre – trained on Natural Language Inference data – learning how sentences relate in meaning.
- In this project, BERT converts sentences into embeddings (dense numerical vectors), and cosine similarity is computed between them to detect meaning – level plagiarism.

3. SBERT Models

- **Problem with BERT:** Comparing sentence pairs using BERT is slow because it must process both sentences together every time.
- **Solution:**
 - SBERT modifies BERT by adding a pooling layer that averages or compresses token embeddings into a single fixed – size sentence vector.
 - This makes it much faster for comparing many pairs of sentences.
- **How it works:**
 - Each sentence is passed independently through SBERT to get its embedding.
 - These embeddings are then compared using cosine similarity – no need to re – run BERT for every pair.

- **Fine – tuning:** During training, SBERT learns to place semantically similar sentences close together in the embedding space and dissimilar ones far apart.
- **Variants Used:**
 - **All – MiniLM - L6 - v2:** Lightweight, faster, good for real-time detection.
 - **All – mpnet – base - v2:** Larger, captures deeper semantic nuances, more accurate for complex paraphrases.

10. EVALUATION METRICS

COMPARATIVE ANALYSIS OF THE MODELS AND THE BASELINES

To evaluate the effectiveness of the proposed hybrid approach, multiple models were trained and compared on the MIT Plagiarism Dataset. The baseline method (TF – IDF + Cosine Similarity) served as a fast and simple word-level comparison model, while transformer – based models were used for semantic similarity detection.

Model	Type	Accuracy	Precision	Recall	F1 Score	Remarks
TF – IDF + Cosine Similarity	Baseline (Lexical)	0.78	0.78	0.79	0.78	Fast but limited to word – level matching
BERT – base – NLI – mean tokens	Deep Learning	0.80	0.79	0.81	0.80	Captures semantic meaning but computational heavy

SBERT all – MiniLM - L6 - v2	Transformer (Lightweight)	0.822	0.799	0.867	0.832	Efficient and accurate; best speed – performance tradeoff
SBERT all – mpnet – base - v2	Transformer (Advanced)	0.829	0.813	0.862	0.836	Strongest semantic understanding; slightly higher compute time

11. OBSERVATION

- The baseline TF – IDF + Cosine method performs adequately for direct copy detection but fails in paraphrased or meaning – level similarity.
- BERT – based models significantly improve semantic detection but are computationally intensive.
- SBERT variants achieve the best overall balance between performance, accuracy, and efficiency.
- Among all, SBERT all – mpnet – base - v2 yielded the highest accuracy and F1 – score, proving to be the most effective model for academic plagiarism detection.

12. TOOLS (FRAMEWORK ENVIRONMENT)

Environment & Frameworks:

- Google Colab with GPU support – for model training
- Python 3.12
- Microsoft Visual Studio Code – user interface development
- Google Drive – store datasets and model checkpoint epochs, store files
- Hugging Face Transformers

- WandbAI (Weights and Baises) – Used to track training progress, monitor metrics and visualize learning curves over epochs.

Algorithms & Techniques:

- TF – IDF with Cosine Similarity for lexical similarity.
- BERT – base – NLI – mean – tokens for semantic understanding.
- SBERT models for efficient embedding generation.
- Text Preprocessing
- Data Splitting – “Train – Validation – Test” split for training and evaluation.
- Cosine Similarity Loss – loss function to optimize embedding closeness for pairs judges similar.

Libraries & Packages:

- NLTK (Natural Language Tool Kit) for text preprocessing.
- SpellChecker for spelling correction.
- Scikit – learn for TF – IDF, cosine similarity, and evaluation metrics.
- Sentence – Transformers for model loading, fine – tuning and encoding.
- Pandas and NumPy for data handling and computation.
- Torch.nn.functional provides cosine similarity function.
- Contractions for text normalization.
- PyTorch for dataloader model training.
- RE – Regular expressions for pattern matching in text preprocessing for noise removal.
- LoggingHandler and DataLoader for model training and batch loading.

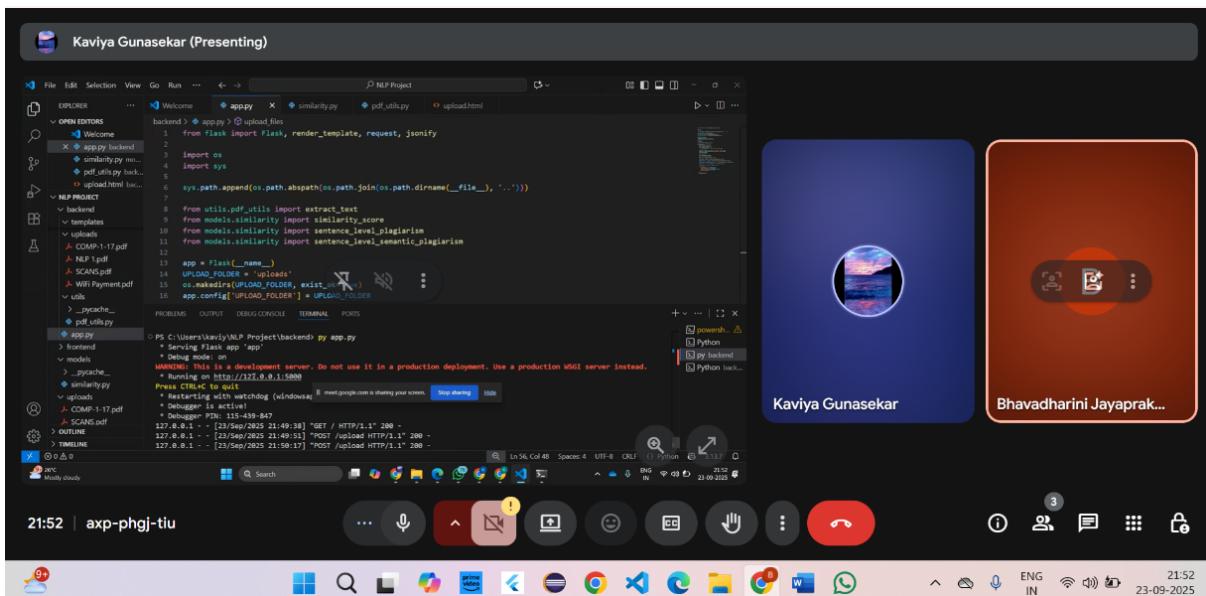
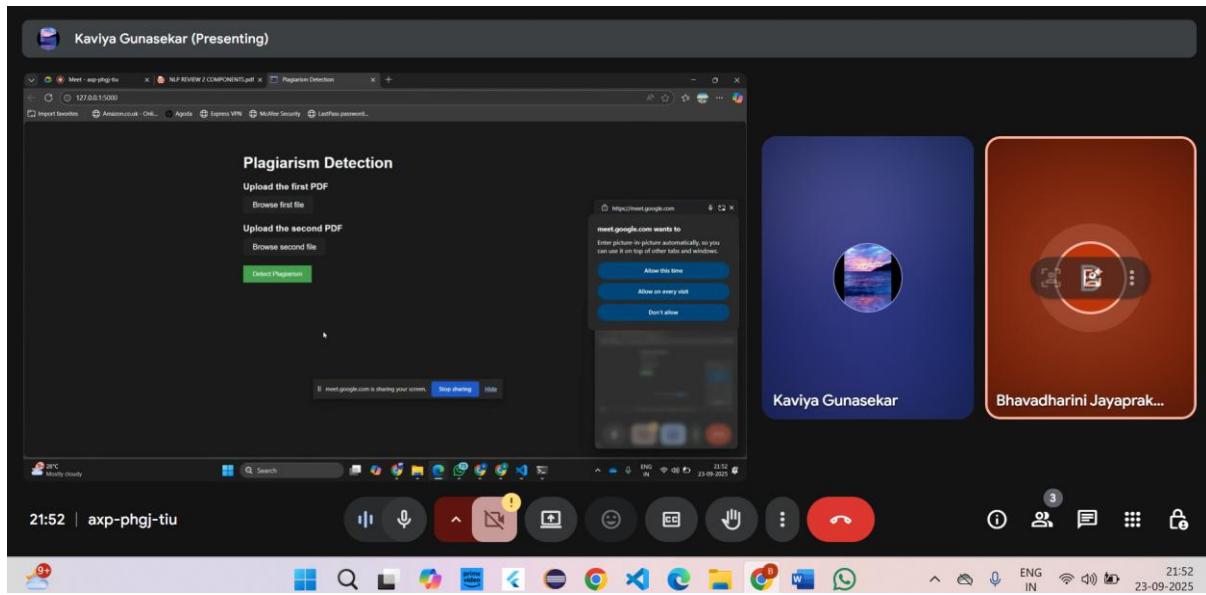
13. CONCLUSION

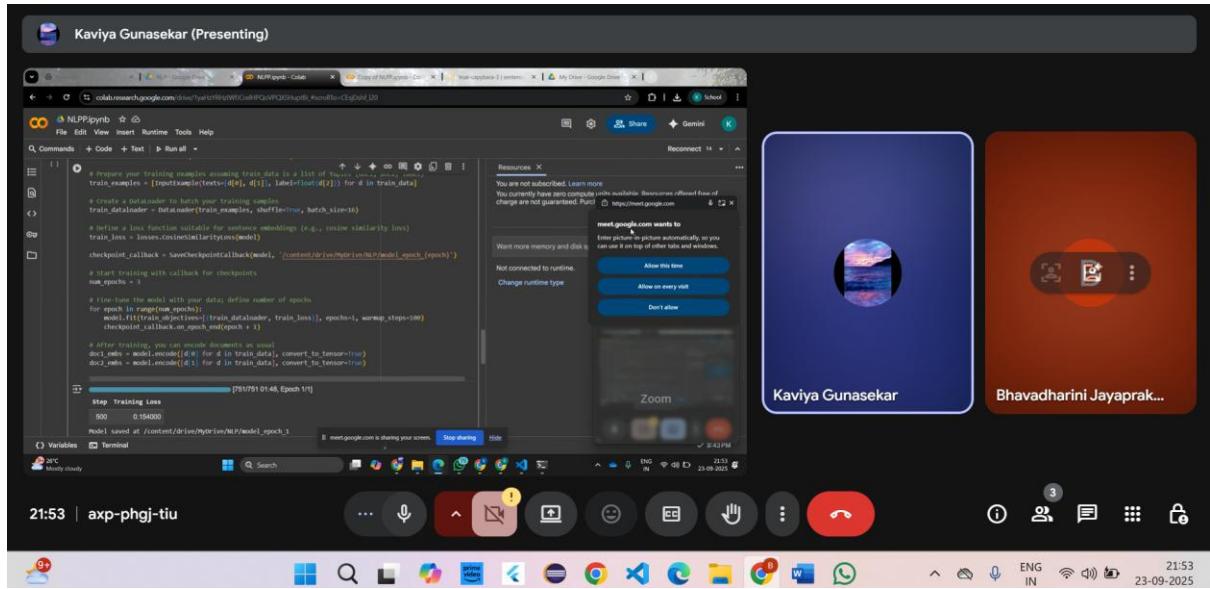
This project successfully developed a Document Similarity Checker for Academic Writing that integrates both traditional NLP techniques and transformer – based deep learning models to detect plagiarism more effectively. The hybrid system first uses TF – IDF with Cosine Similarity for fast and accurate detection of lexical overlap, followed by BERT and SBERT variants to capture deeper semantic and contextual similarities between texts. Trained on the MIT Plagiarism Dataset, the models demonstrated strong performance with high accuracy, precision, recall, and F1 scores, proving the system's ability to detect both direct copying and paraphrased plagiarism. By combining speed, accuracy, and semantic understanding, this approach provides a reliable and scalable solution for academic plagiarism detection, bridging the gap between traditional and modern NLP – based systems.

14. FUTURE WORK

- Extend the model to support multilingual plagiarism detection and cross – language similarity.
- Implement chunk – based document comparison to improve handling of very long texts such as theses or reports.
- Integrate visual highlighting of matched sentences or paragraphs to make results more interpretable for users.
- Explore lightweight transformer models (DistilBERT, TinyBERT) for faster real-time analysis with minimal resource usage.
- Combine Word2Vec or Doc2Vec embeddings with SBERT to enhance representation diversity.
- Build an interactive web – based interface or dashboard that allows users to upload documents and view similarity scores easily.
- Experiment with cross – domain datasets (news, legal, or medical texts) to evaluate generalization beyond academic writing.

15. EVIDENCE OF COLLABOARTION





16. REFERENCES

- [1] N. Gahman and V. Elangovan, “A Comparison of Document Similarity Algorithms,” *arXiv preprint*, 2023.
- [2] R. Ravinder, C. S. Onses, T. H. Dang, and J. R. Hershberger, “A Comparison of Vector-based Approaches for Document Similarity Using the RELISH Corpus,” in *CEUR Workshop Proceedings*, 2023.
- [3] S. V. Moravvej, H. R. Arabnia, and A. G. Bouris, “A Novel Plagiarism Detection Approach Combining BERT-based Embedding, Attention-LSTMs and Differential Evolution Algorithm,” *Proc. IEEE/ACM International Conference on Machine Learning Applications*, 2023.
- [4] J. Xian, Y. Sun, M. Liu, and H. Zhang, “BERT-Enhanced Retrieval Tool for Homework Plagiarism Detection System,” *IEEE Access*, 2024.
- [5] K. Kalyani, S. Mishra, and P. Singh, “PlagCheck: An Efficient Way to Identify Plagiarism Using BERT,” in *Proc. IEEE International Conference on Intelligent Computing*, 2023.
- [6] D. M. Setu, R. Ahmed, and T. R. Islam, “A Comprehensive Strategy for Identifying Plagiarism in Academic Submissions,” *International Journal of Computer Applications*, vol. 178, no. 50, pp. 1–10, 2025.

- [7] K. Vyas, A. Gupta, and R. Sharma, “Semantic Similarity and Plagiarism Checker,” in *Proc. IEEE International Conference on Smart Computing (SMARTCOM)*, 2023.
- [8] K. S. Babu, M. R. Prasad, and R. Kumar, “An Interactive Plagiarism Checker with Text Processing and Similarity Analysis,” *International Journal of Advanced Computer Science*, vol. 15, no. 4, pp. 512–520, 2024.
- [9] A. W. Qurashi, A. R. Al-Sakkaf, and H. Alshammary, “Document Processing Methods for Semantic Text Similarity Analysis,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 245–252, 2020.
- [10] S. Sahani and R. K. Mohapatra, “Smart Cloud Document Clustering and Plagiarism Checker Using TF-IDF and Cosine Similarity,” *International Journal of Computer Applications*, vol. 166, no. 3, pp. 7–12, 2017.
- [11] P. Coates and F. Breitinger, “Identifying Document Similarity Using a Fast Estimation of the Levenshtein Distance Based on Compression and Signatures,” *Journal of Digital Forensics, Security and Law*, vol. 17, no. 1, pp. 32–49, 2022.
- [12] K. Baba, T. Nakatoh, and T. Minami, “Plagiarism Detection Using Document Similarity Based on Distributed Representation,” in *Proc. IEEE International Conference on Advanced Informatics: Concepts, Theory And Applications*, 2017.
- [13] Nurhayati and Busman, “Development of Document Plagiarism Detection Software Using Levenshtein Distance Algorithm on Android Smartphone,” *International Journal of Computer Applications*, vol. 166, no. 9, pp. 1–7, 2017.
- [14] A. S. Hussein, “Arabic Document Similarity Analysis Using N-Grams and Singular Value Decomposition,” *Journal of Computer Science*, vol. 11, no. 6, pp. 851–860, 2015.
- [15] M. Sangeetha, P. Keerthika, K. Devendran, S. Sridhar, S. S. Raagav, and T. Vigneshwar, “Compute Query and Document Similarity Using Explicit Semantic Analysis,” in *Proc. IEEE International Conference on Computer Communication and Informatics (ICCCI)*, 2022.
- [16] R. Lackes, J. Bartels, E. Berndt, and E. Frank, “A Word-Frequency Based Method for Detecting Plagiarism in Documents,” in *Proc. 42nd Hawaii International Conference on System Sciences*, 2009.
- [17] R. T. Hassan and N. S. Ahmed, “Evaluating of Efficacy Semantic Similarity Methods for Comparison of Academic Thesis and Dissertation Texts,”

International Journal of Applied Engineering Research, vol. 18, no. 3, pp. 45–56, 2023.

- [18] J. E. Alvarez, “A Review of Word Embedding and Document Similarity Algorithms Applied to Academic Text,” *arXiv preprint*, 2017.
- [19] L. P. Sari, “Cosine Similarity-based Plagiarism Detection on Electronic Documents,” *International Journal of Emerging Technologies in Learning*, vol. 18, no. 1, pp. 112–120, 2023.
- [20] A. Patel, K. Bakhtiyari, and M. Taghavi, “Evaluation of Cheating Detection Methods in Academic Writings,” *Procedia Computer Science*, vol. 3, pp. 504–508, 2011.