Title: Semantic Document Originality Analysis Using Sentence-BERT and Multi-Dimensional NLP Metrics

Traditional plagiarism detection techniques rely on lexical overlap and exact string matching, making them ineffective against paraphrased and semantically rewritten content. This work proposes an AI-driven document originality analysis framework that leverages Sentence-BERT (SBERT) for sentence-level semantic similarity detection.

Each document is decomposed into sentences and encoded using SBERT embeddings, enabling contextual semantic representation beyond surface-level text similarity. Multiple originality dimensions—lexical diversity, semantic uniqueness, structural variation, source independence, and paraphrase depth—are quantified using embedding-based similarity measures and statistical linguistic features. These metrics are aggregated into a unified originality score that reflects both semantic novelty and structural independence.

The system presents interpretable results through an interactive dashboard comprising gauge, radar, and bar visualizations, along with sentence-level similarity explanations. The proposed framework offers a robust, explainable, and scalable solution for academic originality assessment and AI-assisted content evaluation.

METHODOLOGY

1. Text Extraction and Preprocessing

Uploaded documents (PDF/TXT) are parsed to extract raw text, which is segmented into individual sentences. Minimal preprocessing—such as normalization and whitespace cleanup—is applied to preserve semantic integrity, as excessive preprocessing may degrade embedding quality.

2. Sentence Embedding Using Sentence-BERT (SBERT)

Each sentence is encoded into a dense vector representation using Sentence-BERT (SBERT), specifically the pretrained model:

all-MiniLM-L6-v2

This model maps sentences into a fixed-dimensional embedding space where semantically similar sentences are positioned closer together. Unlike vanilla BERT, SBERT enables efficient sentence-level similarity computation without quadratic complexity.

$$s_i \in \mathbb{R}^d \quad \text{denotes the SBERT embedding of sentence} i$$

denote the SBERT embedding of sentence i.

3. Semantic Similarity Computation

Semantic similarity between two sentences is computed using **cosine similarity** in the embedding space

Sentences whose similarity exceeds a predefined threshold τ\tauτ are flagged as potential originality risks:

$$[\mathrm{Sim}(s_i, s_j) = \frac{s_i \cdot s_j}{\|s_i\| \, \|s_j\|}][\mathrm{Flag}(s_i) = \begin{cases} 1, & \text{if} \max_j \mathrm{Sim}(s_i, r_j) \geq \tau \\ 0, & \text{otherwise} \end{cases}]$$

4. Multi-Dimensional Originality Feature Extraction

The originality of the document is decomposed into the following dimensions:

• Lexical Diversity

Measured using vocabulary distribution statistics to assess word variety and repetition patterns.

• Semantic Uniqueness

Computed as the inverse of the average semantic similarity score across all sentence embeddings:

• Structural Variation

Analyzed using sentence length distribution and syntactic variability across the document.

• Source Independence

Proportion of sentences with similarity scores below the plagiarism threshold τ\tauτ.

• Paraphrase Depth

Derived from high semantic similarity combined with low lexical overlap, indicating rephrased content.

All metrics are normalized to a common scale [0,100][0, 100][0,100].

$$[\text{Semantic Uniqueness} = 1 - \frac{1}{N} \sum_{i=1}^{N} \max_j \mathrm{Sim}(s_i, r_j)]$$

5. Originality Score Aggregation

A weighted aggregation scheme is used to compute the final originality score:

$$[\text{Originality Score} = \sum_{k=1}^{M} w_k \cdot m_k]$$