**Introduction**

The rapid growth of digital technology has revolutionized the way information is created, accessed, and disseminated. This transformation has significantly contributed to academic progress by providing researchers and students with vast amounts of easily accessible knowledge. However, this digital expansion has also increased concerns regarding plagiarism, which poses a substantial threat to the integrity of education and research. Plagiarism, broadly defined as the act of presenting someone else's work, ideas, or expressions as one's own without proper attribution, has become more prevalent due to the widespread availability of online content and the increasing sophistication of text manipulation tools [1].

Plagiarism manifests in various forms, including direct copying, paraphrasing, self-plagiarism, and cross-lingual plagiarism. Direct copying involves duplicating large portions of text verbatim without citation, while paraphrasing plagiarism entails modifying sentence structures or replacing words with synonyms while retaining the original meaning. Cross-lingual plagiarism, on the other hand, occurs when content is translated from one language to another without proper acknowledgment [3]. The advent of online academic databases, open-access journals, and AI-based text generation tools has made it easier for individuals to commit such unethical practices, making plagiarism detection a challenging task for educators and institutions [7]. Given the increasing ease of text duplication, there is a pressing need for efficient, scalable, and automated plagiarism detection systems to ensure academic integrity and maintain trust in scholarly contributions.

**Challenges in Traditional Plagiarism Detection Approaches**

Traditional plagiarism detection methods primarily rely on text-matching techniques, which compare documents based on syntactic and lexical similarities. One of the most widely used techniques is the Vector Space Model (VSM), which represents documents as vectors in a multi-dimensional space and measures their similarity using cosine similarity scores [3]. While these methods effectively detect direct copying, they often struggle with identifying paraphrased, structurally modified, and semantically altered content. Additionally, simple text-matching tools face challenges in detecting cross-lingual plagiarism, where the original content is translated into another language while maintaining its core meaning [6].

To address these limitations, researchers have developed semantic similarity models that analyze the meaning of words rather than just their lexical appearance. Early approaches, such as WordNet-based similarity measures and Latent Semantic Analysis (LSA), attempted to capture contextual relationships between terms. However, these models required extensive linguistic resources and had limited scalability. Moreover, manually tuning similarity thresholds for different types of content remained a significant challenge, making traditional methods inefficient in handling large datasets with complex plagiarism patterns [5].

**Advancements in AI-Based Plagiarism Detection**

The integration of machine learning and deep learning has significantly improved plagiarism detection by enabling sophisticated text analysis beyond basic lexical matching. Traditional approaches, such as direct text comparison, often fail to identify paraphrased, restructured, or semantically altered content. In contrast, supervised classification models trained on large datasets can effectively distinguish between original and plagiarized text by analyzing sentence structure, word embeddings, and contextual meaning [9]. Among deep learning techniques, Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory (LSTM) networks have significantly enhanced detection accuracy by understanding semantic relationships between words and phrases rather than relying on exact word matches. Additionally, attention-based neural networks have shown superior performance in identifying indirect plagiarism, where the sentence structure is altered while preserving the original meaning [9].

Several AI-powered plagiarism detection systems have been developed to improve academic integrity verification. DOCODE 3.0, for instance, applies an information fusion process from multiple documental data sources, providing a comprehensive framework for detecting both direct and complex forms of plagiarism [2]. Similarly, BERT-enhanced retrieval tools leverage pre-trained language models for more accurate text similarity analysis in academic settings [10]. Beyond text-based detection, behavioral analysis techniques such as student activity tracking models have been explored to analyze writing patterns, keystroke dynamics, and revision histories to detect anomalies in student submissions. This approach is particularly effective in identifying contract cheating, where students submit work they did not create. By combining linguistic analysis, statistical techniques, deep learning models, and behavioral tracking, modern plagiarism detection systems provide a multi-faceted and more reliable approach to preserving academic integrity.

**Proposed AI-Powered Plagiarism Detection Framework**

This study introduces Nirikṣaṇam, an AI-driven plagiarism detection framework that integrates multiple advanced techniques to enhance detection accuracy. Unlike conventional plagiarism detection tools that rely solely on text-matching algorithms, Nirikṣaṇam incorporates a hybrid approach combining keyword extraction, web scraping, semantic similarity analysis, and sequence alignment techniques.

To improve the accuracy of plagiarism detection, Nirikṣaṇam employs the YAKE (Yet Another Keyword Extractor) algorithm, which extracts key terms based on statistical properties. This technique enhances focus on significant concepts and themes within a document, enabling more precise similarity detection. Additionally, the system incorporates web scraping techniques using Beautiful Soup, a Python-based library that retrieves and parses online content from publicly available sources. By expanding its search scope beyond local databases, the system improves its ability to detect potential content overlaps with external references.

Furthermore, Nirikṣaṇam leverages the Universal Sentence Encoder (USE) to generate high-dimensional vector representations of sentences, facilitating semantic similarity analysis even for paraphrased content. Unlike conventional plagiarism detectors that focus on word-level comparisons, USE-based models analyze the overall context of sentences, improving detection accuracy for conceptually similar yet lexically distinct text. Lastly, the SequenceMatcher algorithm is incorporated to compare structural patterns between documents, ensuring the identification of direct matches and obfuscated content based on sequence alignment techniques.