Robust Video Plagiarism Detection Using Word Embeddings from Audio Transcriptions

Lucas Leonardo Barros Silva[1], **Paulo de Araujo Freitas-Filho**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Caixa Postal 7851 – 50732-970 – Recife – PE – Brasil

Abstract. *Video piracy presents a significant challenge in the digital era, requiring effective detection methods to protect intellectual property. This paper proposes a novel approach for detecting video plagiarism by leveraging word embeddings derived from audio transcriptions. Our method begins by extracting audio streams from videos and transcribing the audio content. We then generate semantic embeddings, storing these embeddings in a vector store for efficient similarity searches. To identify potential plagiarism, query videos are processed through the same pipeline, and their embeddings are compared against reference embeddings. A Euclidean distance below a predefined threshold indicates possible plagiarism, enabling accurate classification and identification of plagiarized videos. Experimental evaluations demonstrate the method's scalability and efficiency, particularly in detecting complete video copies with explicit English speech content. This approach offers a robust and scalable solution against joint video manipulations, providing a practical framework for combating video piracy in large-scale content environments.*

Keywords: Video plagiarism detection, word embeddings, FAISS, Whisper model, sentence transformers, audio transcription.

1. Introduction

The rapid proliferation of online video-sharing platforms such as YouTube and Twitch has enabled users worldwide to broadcast and consume live video content [Martemucci and Swerdlow 2017]. While this democratization of content creation and distribution has numerous benefits, it has also led to significant intellectual property rights and copyright infringement challenges. Users can stream and watch copyrighted live events, such as sports matches and television shows, without authorization from content owners [Zhang et al. 2018a], resulting in substantial financial losses for content creators and rights holders.

1.1. Motivation

Video-sharing platforms have implemented various measures to detect and prevent copyright infringement. For instance, YouTube's Content ID system compares uploaded videos against a database of copyrighted material to identify unauthorized content [King 2007]. However, Content ID and similar systems face critical limitations. They are less effective for live video streams, as these are generated and consumed in real-time, making it challenging to maintain an up-to-date database of copyrighted live content [Zhang et al. 2018b]. Moreover, sophisticated infringers employ video rotation,

cropping, re-encoding, and audio transformations to manipulate content and evade detection mechanisms [Esmaeili et al. 2011]. These alterations often break fingerprinting and watermarking algorithms traditionally used for content protection [Barg et al. 2003, Podilchuk and Zeng 1998].

(a) (b)

Figure 1: Examples of manipulated videos that bypass conventional detection systems [Zhang et al. 2018b]. Such manipulations present significant hurdles for automated detection methods, necessitating the development of more robust and resilient approaches.

1.2. Problem Statement

Existing copyright protection techniques, such as fingerprinting [Barg et al. 2003] and watermarking [Podilchuk and Zeng 1998], are primarily designed for static content like digital music, software, or ebooks. These methods are not directly applicable to live video streams generated and consumed in real-time [Esmaeili et al. 2011]. Moreover, they are vulnerable to joint content manipulations that degrade their effectiveness, as illustrated in Figure 1.

Alternative approaches focus on detecting video similarities using comprehensive feature-based methods [Nie et al. 2017], but they often struggle with videos intentionally altered to appear different from the original content. As a result, there is a pressing need for robust plagiarism detection techniques that can handle various content modifications without significant performance loss.

1.3. Related Works

The detection of video plagiarism has garnered significant attention in recent years, driven by the proliferation of online video-sharing platforms and the consequent rise in unauthorized content distribution. Existing methodologies for plagiarism detection can be broadly categorized into text-based, visual-based, and hybrid approaches, each leveraging different aspects of video content to identify instances of plagiarism.

1.3.1. Text-Based Plagiarism Detection

Textual plagiarism detection has evolved considerably with the advent of natural language processing (NLP) techniques, mainly through word embeddings. Asghari [Asghari et al. 2019] explored cross-language plagiarism detection by employing word embedding methods, which map words into continuous vector spaces, capturing semantic similarities across languages. This approach enhances the detection of paraphrased

content by identifying semantically similar phrases even when surface-level alterations are present. Similarly, Saeed and Taqa [Saeed and Taqa 2022] introduced a deep learning framework utilizing Word2Vec and GloVe embeddings within a Siamese LSTM network to compute similarity features between texts. Their method demonstrated high Precision and Recall on datasets such as PAN-PC-11, indicating the potential applicability of embedding-based models for detecting external plagiarism.

Yalcin et al. [Yalcin et al. 2022] further advanced text-based plagiarism detection by integrating part-of-speech (POS) tag n-grams with Word2Vec embeddings. Their system effectively captured syntactic and semantic similarities, achieving competitive results in detecting high obfuscation paraphrasing. These studies collectively underscore the efficacy of word embeddings in enhancing the semantic understanding required for accurate plagiarism detection. However, their focus remains predominantly on text, with limited exploration into audio-based or multimedia content.

1.3.2. Visual-Based Plagiarism Detection

In parallel, visual-based plagiarism detection methods have been developed to identify copied video content through image and frame analysis. Wary and Neelima [Wary and Neelima 2018] conducted a comprehensive review of robust video copy detection techniques, emphasizing visual hashing methods to generate perceptual hash codes resistant to joint video manipulations such as rotation, scaling, and compression. Their analysis highlighted the strengths of visual hashing in maintaining resilience against various distortions yet pointed out the challenges in handling real-time video streams and extensive video databases.

Despite their robustness, visual-based methods often struggle with videos that undergo sophisticated transformations designed to evade detection, such as temporal cropping or audio-visual synchronization alterations. These limitations necessitate exploring complementary approaches that can address the semantic aspects of video content beyond mere visual similarity.

1.3.3. Hybrid Approaches and Semantic Analysis

Recognizing the limitations of purely text-based or visual-based methods, recent research has ventured into hybrid approaches that integrate both modalities to enhance detection accuracy. El-Rashidy et al. [El-Rashidy et al. 2023] proposed a system combining feature selection and Support Vector Machines (SVM) to detect lexical, syntactic, and semantic plagiarism in text. While their methodology effectively identifies plagiarized content through semantic scoring, it does not extend to multimedia content, leaving a gap for future exploration in video plagiarism detection.

Integrating semantic analysis through word embeddings offers a promising avenue to bridge this gap. By converting audio transcriptions of video content into text and applying embedding-based similarity measures, semantic plagiarism can be detected even in significant audio and visual manipulations. This approach leverages text-based semantic understanding and visual content analysis strengths, providing a more holistic detection

mechanism.

1.3.4. Limitations of Existing Methods and the Need for Novel Approaches

Table 1: Summary of Related Work in Plagiarism Detection

Study	Approach	Techniques Used	Key Findings
Asghari et al. [Asghari et al. 2019]	Cross-Language Plagiarism Detection	Word Embeddings, Bilingual Corpus	Enhanced detection accuracy across languages
Saeed and Taqa [Saeed and Taqa 2022]	Textual Plagiarism Detection	Word2Vec, GloVe, Siamese LSTM	High precision and recall on textual datasets
Yalcin et al. [Yalcin et al. 2022]	External Plagiarism Detection	POS Tag N-Grams, Word2Vec	Effective in high obfuscation paraphrasing
Wary and Neelima [Wary and Neelima 2018]	Visual-Based Detection	Visual Hashing	Robust against video distortions like rotation and scaling
El-Rashidy et al. [El-Rashidy et al. 2023]	Textual Plagiarism Detection	Feature Selection, SVM	Superior performance on PAN datasets

Table 1 provide a comparative overview of the discussed studies. Our work distinguishes itself by targeting video plagiarism detection through the innovative use of word embeddings derived from audio transcriptions, thereby addressing the semantic gaps left by existing methodologies.

While existing methods have demonstrated considerable success in their respective domains, they exhibit notable limitations when applied to real-world scenarios involving diverse and manipulated video content. Text-based methods are constrained by their reliance on accurate transcriptions and may falter in multilingual or noisy audio environments. Although robust to certain distortions, visual-based methods often miss the semantic nuances that indicate plagiarism.

Moreover, hybrid approaches that combine text and visual features tend to increase computational complexity, posing scalability and real-time detection challenges. Consequently, there is a pressing need for innovative methodologies that can seamlessly integrate semantic analysis from audio transcriptions with efficient visual content processing to achieve robust and scalable video plagiarism detection.

1.4. Our Contribution

Addressing these challenges, our proposed method leverages word embeddings derived from audio transcriptions to capture the semantic content of videos, thereby enhancing plagiarism detection capabilities beyond traditional visual analysis. By integrating advanced speech recognition and semantic embedding techniques, our approach offers a resilient solution to identify plagiarized content even amidst complex audio and video transformations.

Therefore, while significant advancements have been made in both text-based and visual-based plagiarism detection, the integration of semantic analysis through word embeddings from audio transcriptions presents a novel and promising direction. This approach enhances the detection accuracy by capturing deeper semantic relationships and mitigates the challenges posed by audio and visual content manipulations. Our proposed method builds upon the strengths of existing techniques while addressing their inherent limitations, paving the way for more robust and reliable video plagiarism detection systems.

By focusing on semantic content rather than raw audio or visual features, our method provides a robust solution to the plagiarism detection problem in the face of common video manipulations.

Hence, in summary the main contributions of our work are:

- Introducing an audio-based plagiarism detection system resilient to video and audio transformations that typically break fingerprinting algorithms.
- Utilizing the Whisper speech recognition model for accurate audio transcription extracted from videos, even in the presence of noise or minor alterations.
- Employing word embeddings generated by the Sentence-Transformers library with the BAAI/bge-m3 model to capture the semantic essence of transcribed content.
- Implementing a scalable vector store using FAISS for efficient similarity searches, enabling real-time detection of plagiarized content based on semantic similarity thresholds.

2. Proposed Solution

The proposed framework offers a robust solution for detecting video plagiarism by systematically converting videos to text embeddings and efficiently comparing these embeddings. Each methodology component is designed to handle large datasets and diverse content, making it suitable for practical applications in academic and media industries.

The methodology is structured into four principal steps: data preprocessing, embedding generation, vector store creation, and the plagiarism detection process, as illustrated in Figure 2.

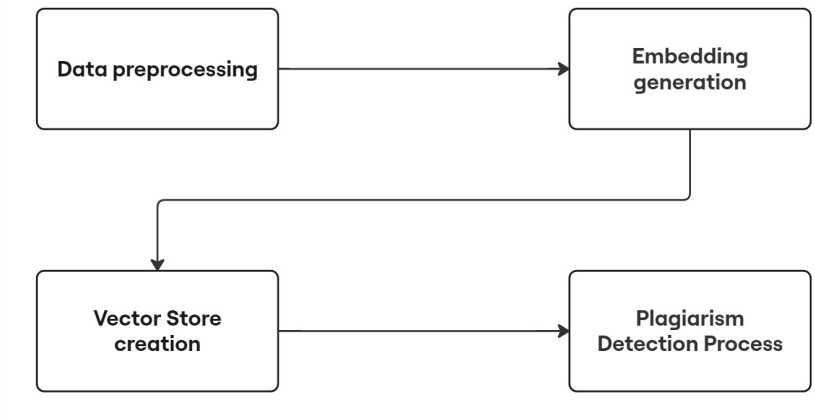

Figure 2: Proposed framework pipeline flowchart

2.1. Data Preprocessing

Data preprocessing is a critical initial step that prepares the video data for analysis by converting it into a consistent and analyzable format. This involves two main processes: extracting audio from video files and transcribing the audio into text.

2.1.1. Video-to-Audio Conversion

Video files are processed to extract their audio streams, effectively isolating the auditory content while discarding the visual component. This conversion ensures that the audio quality is uniform, which is essential for accurate transcription. The resulting audio files

are standardized in format and quality, making them suitable for subsequent speech-to-text processing.

2.1.2. Speech-to-Text Transcription

The extracted audio files are then transcribed into text using a robust speech recognition system capable of handling multiple languages and varying audio qualities. This step transforms the audio data into text, which can be more effectively analyzed using natural language processing techniques. Accurate transcription is vital, as it directly influences the quality of the subsequent semantic embeddings and the overall effectiveness of plagiarism detection.

2.2. Embedding Generation

Embedding generation involves transforming the transcribed textual data into high-dimensional numerical vectors that capture semantic meaning. A multilingual embedding model is employed to generate semantic embeddings that represent the contextual and semantic nuances of the text. These embeddings serve as a numerical representation of the video's content, enabling quantitative comparison between videos based on their transcriptions.

2.3. Vector Store Creation

A specialized vector database is created to facilitate efficient and scalable similarity searches among the high-dimensional embeddings. This database stores the embeddings in an optimized structure that supports rapid similarity queries using metrics such as Euclidean distance or cosine similarity. Organizing the embeddings in this manner allows for quick retrieval and comparison, essential for real-time plagiarism detection in large datasets.

2.4. Plagiarism Detection Process

The plagiarism detection process utilizes the embeddings and vector store to identify potential instances of plagiarism by comparing query videos against reference videos.

2.4.1. Similarity Search

Each query video undergoes the same preprocessing steps as the reference videos to ensure consistency.

The query video's embedding is compared against the reference embeddings stored in the vector database. A similarity search uses an Euclidean distance metric to quantify the closeness between the query embedding and each reference embedding. The system retrieves the most similar embeddings, effectively identifying reference videos that are semantically close to the query video.

2.4.2. Classification Criteria

The final classification of whether a video is plagiarized is determined based on the similarity scores obtained from the similarity search. A predefined threshold θ, established through empirical analysis, determines this. Formally, a query video V_q is classified as plagiarized concerning a reference video V_r if:

$$\text{Distance}(E_q, E_r) < \theta,$$

where E_q and E_r are the embeddings of V_q and V_r, respectively. This criterion ensures that only videos with significant semantic overlap—indicated by a distance below the threshold—are flagged as plagiarized. This approach reduces false positives and enhances the reliability of the plagiarism detection system.

3. Experiments

This section presents a comprehensive evaluation of our proposed video plagiarism detection method. We explain the dataset, detail the experimental setup, and outline the evaluation metrics.

3.1. Dataset Overview

For our experiments, we utilize the **VCDB** dataset [Jiang et al. 2014], a publicly available collection designed for video copy detection research. The dataset is particularly suited for our study as it contains extensive instances of copied video segments with complex transformations, posing significant challenges for detection algorithms.

The VCDB dataset is organized into two main parts:

- **Core Dataset**: This subset comprises 528 videos (approximately 27 hours of content) collected using 28 carefully selected queries covering a wide range of topics such as commercials, movies, music videos, public speeches, and sports. These videos are highly relevant to the queries and contain numerous partial copies.
- **Background Dataset**: Consisting of 100,000 videos sourced from YouTube, this subset serves as a background corpus to simulate a realistic and challenging environment for plagiarism detection, introducing non-copied content that tests the system's ability to avoid false positives.

3.1.1. Data Selection and Preparation

The core dataset was collected using 28 carefully selected queries [Jiang et al. 2014] a query has several videos for the same scene, such as an Obama speech or a part of a movie, we'll call the queries classes. In this study we focus on classes that contain sufficient samples for robust evaluation. The classes in the dataset vary in size, with each class containing between 3 and 43 videos. To ensure a balanced evaluation and meaningful statistical analysis, we apply the following criteria:

- **Class Selection**: We select only the classes with more than 8 videos, resulting in a subset suitable for our analysis.

- **Video Selection**: Within each selected class, we randomly choose 8 videos. Out of these, the two videos with the largest transcriptions (i.e., those containing the most speech content) are designated as **reference videos** and are used to populate the vector store for plagiarism detection.
- **Test Set Formation**: The remaining 6 videos in each class are used as **test videos** to assess the system's ability to detect plagiarism.

  After applying these criteria, we obtain:

- **Number of Classes**: 25 classes meet the selection criteria.
- **Reference Videos**: 25 classes $\times$ 2 videos $=$ 50 reference videos.
- **Test Videos (Copied Content)**: 25 classes $\times$ 6 videos $=$ 150 test videos containing copied content.

3.1.2. Background Dataset (Non-Copied Content)

To evaluate the system's performance against non-copied content and to assess the false positive rate, we include an additional set of videos:

- **Non-Copied Test Videos**: **Non-Copied Test Videos**: We randomly select 150 videos from the Background Dataset that do not overlap with any of the reference or copied test videos.

3.1.3. Dataset Summary

Our experimental dataset is summarized in Table 2. It consists of reference and test sets, each designed for specific roles in the plagiarism detection process.

Table 2: Summary of the Experimental Dataset

Dataset Partition	Description
Reference Set	50 reference videos used to create the vector store embeddings.
Test Set (Copied Content)	150 test videos containing copied content, sourced from the Core Dataset.
Test Set (Non-Copied Content)	150 test videos containing non-copied content, sourced from the Background Dataset.

All videos are processed to extract audio transcriptions, which are then used to calculate word embeddings. The embeddings of the reference videos are stored in a vector database. We compute each test video's embedding and compare it against the embeddings in the vector store to detect potential plagiarism based on Euclidean distance score thresholds ranging from 0.5 to 0.99.

3.1.4. Data Partitioning

Our approach does not require a traditional training phase, as it leverages pre-trained word embeddings and operates in a zero-shot setting. Consequently, there is no need for a training set. The data is partitioned as follows:

- **Reference Set**: Used solely to create vector store embeddings.
- **Test Set**: Used for evaluating the performance of the plagiarism detection system. It includes copied and non-copied videos to assess true positive and false positive rates, respectively.

This partitioning allows us to evaluate the detection algorithm's effectiveness exclusively without the influence of a training process.

3.2. Experimental Design

Six experiments were conducted to evaluate the performance of our proposed method. Five of these experiments used balanced datasets, where the number of copied and non-copied videos was equal. In these five experiments, we tested a variety of similarity thresholds ranging from 0.5 to 0.99 to assess the system's accuracy, precision, recall, and F1 score.

In addition to these five balanced experiments, we conducted one experiment using an unbalanced dataset to simulate a scenario closer to real-world conditions, where the number of non-copied videos far exceeds the number of copied ones. In this unbalanced experiment, we used 990 non-copied videos and 150 copied videos from the previous experiments. This setup was designed to evaluate how well the algorithm performs in an imbalanced setting, where detecting true copies is more challenging due to the overwhelming presence of non-copied content.

The results of all six experiments, including both the balanced and unbalanced datasets, are presented and analyzed in the section 4 section.

3.3. Hardware and Software Configuration

All experiments were conducted on a system with an Intel Core i5 CPU, NVIDIA GeForce RTX 3050 GPU, and 32 GB of RAM

- **Operating System**: Ubuntu 24.04 LTS on WSL2.
- **Audio Extraction**: FFmpeg 6.1.1 [Tomar 2006].
- **Speech Recognition**: OpenAI's Whisper model (small variant) implemented in Python [Radford et al. 2023].
- **Embedding Generation**: Sentence-Transformers library version 3.1.1 with the BAAI/bge-m3 model [Chen et al. 2023].
- **Similarity Search**: FAISS library version 1.7.2 [Johnson et al. 2019].
- **Language Detection**: `langdetect` library [Nakatani 2010].

3.4. Evaluation Metrics

To quantitatively evaluate the performance of our plagiarism detection system, we employed metrics, such as accuracy, precision, recall, and F1-score at various similarity thresholds.

By structuring our dataset in this manner, we aim to provide a comprehensive evaluation of the proposed method's ability to detect video plagiarism in a realistic and challenging environment.

4. Results

In this section, we present an analysis of the performance of our proposed video plagiarism detection system across six experiments. The evaluation focuses on assessing detection accuracy, the impact of similarity thresholds, and the effect of data imbalance on the system's performance.

4.1. Results of experiments with balanced data

Experiments 1 to 5 were conducted using balanced datasets, each comprising 150 plagiarized videos and 150 non-plagiarized videos. The embeddings of the test videos were compared against the reference embeddings stored in the FAISS vector database using the Euclidean distance metric.

Table 3 summarizes the performance metrics at the optimal thresholds for Experiments 1 through 5. The optimal threshold is the similarity threshold at which the system achieves the highest accuracy. This threshold represents the point where the balance between true positives and negatives is maximized, resulting in the most accurate classification of plagiarized and non-plagiarized content. By selecting this threshold, the system is optimized for overall performance based on accuracy, ensuring the most reliable detection results.

Table 3: Performance metrics at optimal thresholds for Experiments 1 to 5 (balanced data).

Experiment	Threshold	Accuracy	Precision	Recall	F1-score
1	80	0.8367	0.9316	0.8165	0.8165
2	80	0.8167	0.8862	0.7267	0.7985
3	80	0.8333	0.9237	0.7267	0.8134
4	81	0.7900	0.9223	0.6333	0.7510
5	81	0.7967	0.9406	0.6333	0.7570

In Experiment 1, at a similarity threshold of 80, the system achieved an accuracy of 83.67%, with a Precision of 93.16% and a Recall of 81.65%. The high Precision indicates that most videos flagged as plagiarized were indeed copies, while the Recall reflects the system's ability to detect a significant proportion of the plagiarized videos.

Similarly, in Experiments 2 to 5, the system consistently demonstrated high Precision and acceptable Recall values at their respective optimal thresholds, indicating reliable detection performance across different balanced datasets.

The confusion matrices for the experiments provide insights into the classification outcomes. Table 4 presents the confusion matrix for Experiment 1 at the optimal threshold.

Table 4: Confusion matrix for Experiment 1 at threshold 80.

Actual \ Predicted	Not Copy	Copy
Not Copy	142	8
Copy	41	109

A low false positive rate is crucial when dealing with plagiarism and copyright detection systems. In real-world applications, where such techniques are used to identify and take down unauthorized copies, mistakenly removing legitimate content can lead to significant issues. Therefore, minimizing false positives is essential to avoid wrongful actions against genuine content creators.

On the other hand, a higher false negative rate is somewhat expected, given the structure of our data. In our approach, embeddings are generated from the entire transcript of the video, which means that when a segment of a video is compared against the full content, the similarity score may be lower. This happens because the number of characters is not normalized across the comparisons, resulting in misclassifications. This limitation suggests room for improvement, such as implementing chunking techniques to break the videos into smaller, more comparable segments, thereby reducing the false negative rate.

In our results, the system correctly classified 142 non-plagiarized videos and 109 plagiarized videos, with 8 false positives and 41 false negatives. This corresponds to a False Positive Rate (FPR) of 5.33% and a False Negative Rate (FNR) of 27.33%. While the FPR remains acceptably low, the relatively high FNR highlights the need for enhancements in our methodology, particularly in handling discrepancies in video segment comparison.

In our experiments, the decrease in accuracy at higher thresholds is notably influenced by the nature of the data used. Specifically, some videos contain very little spoken content in the audio, resulting in transcriptions that are not representative—having few or even no words. These sparse transcriptions lead to less informative embeddings, making it challenging for the system to find correct matches. As the threshold increases, the system requires higher similarity scores to classify a video as plagiarized. Consequently, videos with inadequate transcriptions are more likely to be misclassified as not plagiarized due to their low similarity scores, contributing to a decrease in overall accuracy.

Figure 3 illustrates the variation of Precision, Recall, and F1-score across different similarity thresholds for Experiment 1. As the threshold increases, Precision tends to improve while Recall decreases, reflecting the typical trade-off in classification tasks.

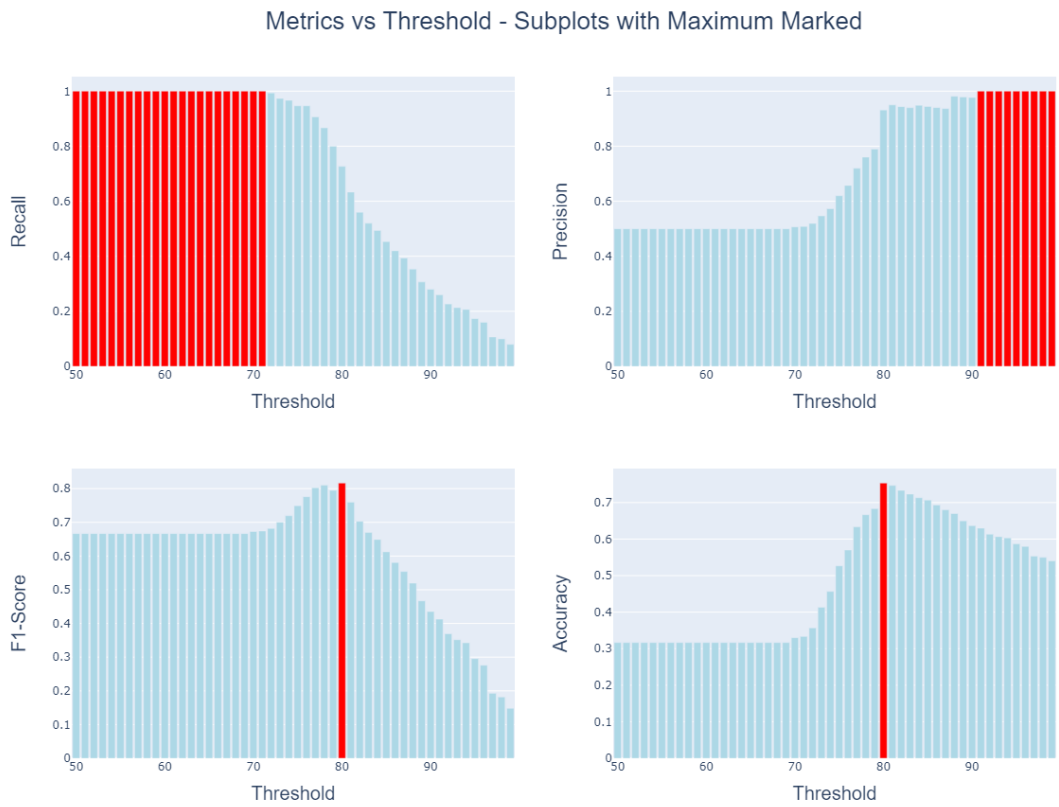

Figure 3: Variation of Accuracy, Precision, Recall, and F1-score with different similarity thresholds in Experiment 1.

In our system, the threshold determines the minimum similarity score required for a video to be classified as plagiarized. As the threshold increases, fewer videos meet this criterion, resulting in more videos being identified as not plagiarized. This shift causes the system to become more conservative, increasing Precision because the likelihood of false positives decreases; only videos with very high similarity scores are flagged as plagiarized. However, this also leads to a decrease in Recall and overall Accuracy.

As the threshold increases, the system may fail to recognize videos that contain plagiarized segments because the overall similarity score does not exceed the higher threshold. This results in an increased number of false negatives—plagiarized videos incorrectly classified as not plagiarized—which lowers both Recall and Accuracy.

At lower thresholds, the system is more lenient, classifying more videos as plagiarized. This increases Recall because more actual plagiarized videos are correctly identified. However, it also raises the number of false positives—non-plagiarized videos incorrectly classified as plagiarized—reducing Precision.

4.2. Results of experiment with imbalanced data

Experiment 6 used an imbalanced dataset comprising 990 non-plagiarized videos and 150 plagiarized videos, simulating a real-world scenario where plagiarized content is less frequent.

At the optimal threshold of 84, the system achieved an overall accuracy of 91.40%. However, due to the imbalance in the dataset, accuracy alone does not provide a complete picture of the system's performance. Table 5 summarizes the performance metrics for Experiment 6.

Table 5: Performance metrics for Experiment 6 (imbalanced data) at threshold 80.

Threshold	Accuracy	Precision	Recall	F1-score
84	0.9140	0.7708	0.4933	0.6016

Although the overall accuracy is high, this is largely due to the high number of correctly classified non-plagiarized videos which dominate the dataset. The Precision and Recall for the "Copy" class are notably lower than the balanced experiments.

The confusion matrix for Experiment 6 at the optimal threshold is shown in Table 6.

Table 6: Confusion matrix for Experiment 6 at threshold 84.

Actual \ Predicted	Not Copy	Copy
Not Copy	968	22
Copy	76	74

The system correctly classified 968 non-plagiarized videos and 74 plagiarized videos. There were 22 false positives and 76 false negatives. The high number of false negatives indicates that the system missed a significant portion of the plagiarized videos.

In imbalanced datasets, accuracy can be misleading because the majority class may dominate it. The imbalance affects the threshold selection based on accuracy since the model does not learn from the data (no training is involved). The threshold determined by maximizing accuracy may not be optimal for detecting the minority class (plagiarized videos), leading to lower Precision and Recall.

Figure 4 illustrates the variation of Precision, Recall, and F1-score across different similarity thresholds for Experiment 6. The Precision and Recall for the "Copy" class are significantly lower than the balanced experiments, highlighting the challenges in detecting plagiarized content when it is underrepresented.

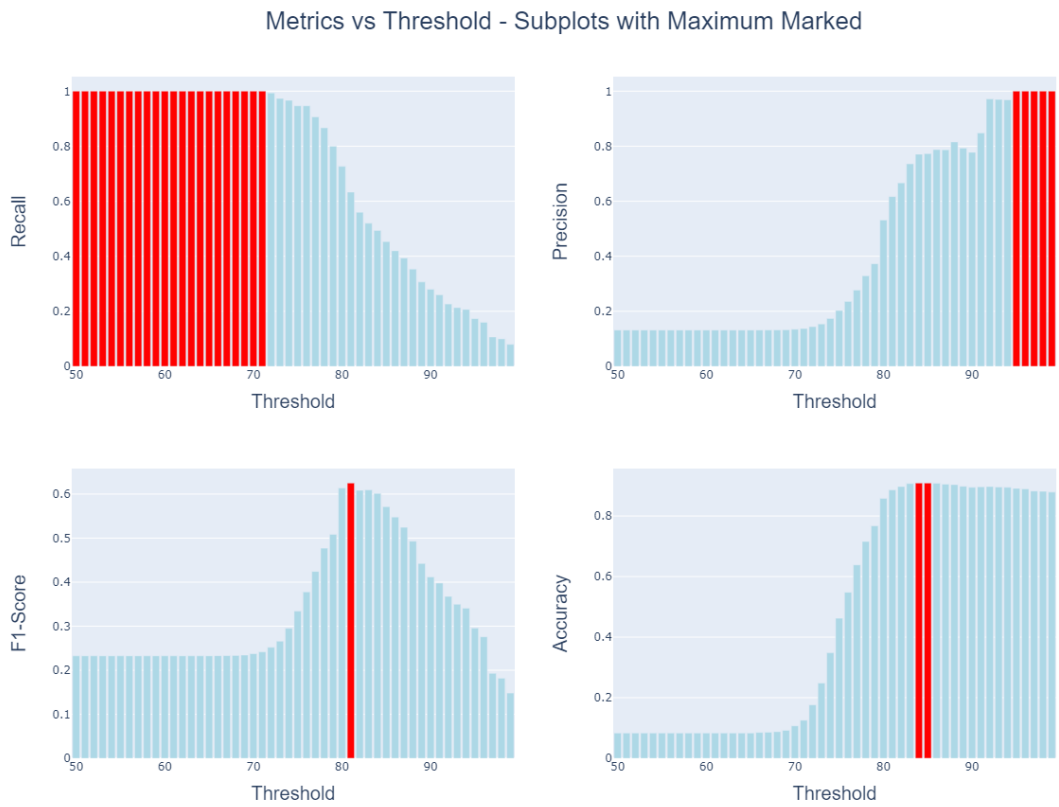

Figure 4: Variation of Accuracy, Precision, Recall, and F1-score with different similarity thresholds in Experiment 6.

Due to the imbalance, the threshold that maximizes overall accuracy may not effectively separate plagiarized and non-plagiarized videos, resulting in lower detection rates for the minority class.

At the optimal threshold in Experiment 6, the False Positive Rate was 2.22%, and the False Negative Rate was 50.67%. The low FPR reflects the system's effectiveness in correctly identifying non-plagiarized videos, but the high FNR indicates that it missed over half of the plagiarized videos.

4.2.1. Analysis of Threshold Selection

The selection of an optimal threshold based solely on maximizing overall accuracy can be misleading, especially in imbalanced datasets, as seen in our experiments. Figure 5 demonstrates how varying the threshold values affects the Precision and Recall for the "Copy" class in Experiment 6.

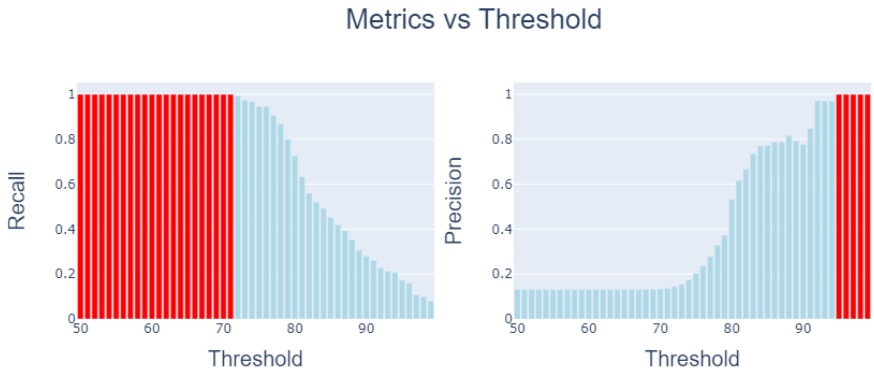

Figure 5: Impact of threshold selection on Precision and Recall for the "Copy" class in Experiment 6.

Our results indicate that while maximizing overall accuracy may provide a reasonable balance across classes, this approach may not be optimal for cases where the "Copy" class (representing plagiarized content) is of primary concern. By fine-tuning the threshold to prioritize Precision for the "Copy" class, we can effectively eliminate false positives, ensuring that all flagged instances are genuinely plagiarized. This approach is particularly valuable in real-world applications where the consequences of false positives—such as mistakenly taking down legitimate content—are severe. Achieving a Precision of 100% for the "Copy" class means the system can confidently act on detected cases, as there will be no erroneous takedowns of non-infringing content.

However, this comes with a trade-off. As Figure 5 illustrates, adjusting the threshold to maximize Precision leads to a reduction in Recall, meaning that fewer true positive cases are detected. While this trade-off reduces the overall effectiveness of detecting all plagiarized content, it guarantees that actions taken on detected content are accurate. In some contexts, such as legal enforcement or content moderation, this balance may be preferable, as it allows for a more assertive and problem-free enforcement of copyright protection.

Therefore, threshold selection should be guided not only by accuracy but by the specific objectives of the system. In cases where Precision for the "Copy" class is crucial, defining the threshold to achieve this ensures a highly reliable detection system, albeit at the cost of potentially missing some plagiarized content. This highlights the importance of aligning threshold selection with the system's real-world use case and specific performance requirements.

5. Discussion

Our proposed video plagiarism detection system, which leverages word embeddings derived from audio transcriptions, demonstrates both robust performance and certain limitations across various experimental settings. The system consistently achieves high precision and acceptable recall rates in balanced datasets, indicating its effectiveness in identifying plagiarized content. However, in imbalanced datasets, such as Experiment 6, overall accuracy remains high primarily due to the predominance of non-plagiarized videos, while the detection of plagiarized videos diminishes. This shortcoming arises not from

model bias but the threshold selection process, which is tuned for maximizing overall accuracy, thus favoring the majority class.

A key factor affecting system performance is the nature of the videos within the dataset. Many of the videos belong to categories such as sports, which naturally contain fewer spoken words or have segments with no dialogue. These characteristics directly impact the system's ability to extract meaningful semantic content from the audio transcriptions, reducing the effectiveness of the plagiarism detection model. The lack of sufficient verbal content and silent cuts diminishes the model's ability to generate accurate embeddings for comparison, especially in detecting partial or subtle plagiarism instances.

To address this challenge in imbalanced data and content-sparse videos, several improvements can be considered:

- **Threshold Selection Based on F1-score:** Prioritizing the minority class by selecting thresholds that optimize F1-score could enhance the detection of plagiarized videos.
- **Class-specific Thresholds:** Implementing separate thresholds for each class may better balance the trade-offs between false positives and false negatives.
- **Cost-sensitive Decision Making:** Incorporating the cost of false negatives, such as undetected plagiarism, into the decision-making process could improve the system's sensitivity to minority class instances.
- **Adjusting Similarity Metrics:** Exploring alternative similarity measures or weighting schemes could also improve class discrimination, particularly in detecting plagiarized content.

Compared to visually-based approaches, our method shows clear divergence, focusing on semantic analysis rather than visual feature extraction. Visually-based methods, such as those employed by Kordopatis-Zilos et al. [Kordopatis-Zilos et al. 2023] and Jiang et al. [Jiang and Wang 2016], are adept at capturing temporal continuity and spatial features, which are critical for detecting both full and partial video plagiarism. These methods typically rely on specialized metrics like mean Average Precision (mAP) and segment-level precision (SP) to provide a more granular assessment of video similarity. By contrast, our system, which relies solely on textual embeddings from audio, employs traditional classification metrics such as Precision, Recall, and F1-score. While these metrics effectively identify plagiarized videos in a binary classification context, they lack the granularity to detect partial copies or account for temporal continuity.

A notable limitation of our approach is its reliance on entire audio transcriptions, which can obscure localized similarities in partial video copies. This limitation could result in the system overlooking subtle instances of plagiarism that do not significantly alter the overall semantic embedding. This challenge is exacerbated in categories like sports, where large portions of the video may contain minimal dialogue or silent segments. As a result, the embeddings may fail to capture the diversity of the video content, leading to less effective plagiarism detection.

Future work will explore strategies to address these limitations, particularly through enhanced chunking methods. By segmenting audio transcriptions into smaller, overlapping chunks, we aim to capture more localized similarities, enabling the detection of partial copies through finer-grained similarity assessments. Furthermore, a detailed

study of different chunking strategies could help better represent the video content in distinct segments, increasing the distance between different contents in the vector database, and improving the discrimination power of the model. Techniques such as sliding window analysis or attention mechanisms may further enhance the granularity and accuracy of the plagiarism detection process, particularly for videos with sparse or unevenly distributed audio content.

Overall, the results indicate the potential of using word embeddings from audio transcriptions for video plagiarism detection. While the system demonstrates robustness in balanced datasets, addressing the challenges of imbalanced data and improving the detection of partial copies remain key areas for future enhancement.

6. Conclusion

In this paper, we presented a novel approach for video plagiarism detection that leverages word embeddings derived from audio transcriptions. Our method involves converting the audio content of videos into text and utilizing advanced semantic embedding techniques to identify plagiarized content based on semantic similarities. Our system demonstrated robust performance through comprehensive experiments conducted on the VCDB core dataset, achieving high precision and recall rates, particularly in videos with clear and English-language audio transcriptions.

Despite these promising results, our approach faces certain limitations. Specifically, it struggles to detect plagiarism that occurs solely within the visual domain or in videos lacking audio content, where no transcriptions are available for analysis. To address these challenges, future work will focus on developing an integrated solution that combines textual embeddings from audio transcriptions with image embeddings derived from visual content analysis. This hybrid approach aims to provide a more comprehensive and resilient plagiarism detection framework capable of handling a wider range of content modifications and ensuring higher accuracy across diverse video formats.

In conclusion, our proposed method significantly advances video plagiarism detection. By harnessing the semantic depth of audio transcriptions, it offers improved accuracy and resilience. The anticipated integration of image-based embeddings will further augment the system's capabilities, paving the way for more robust and scalable solutions in the ongoing battle against video plagiarism.

References

Asghari, H., Fatemi, O., Mohtaj, S., Faili, H., and Rosso, P. (2019). On the use of word embedding for cross language plagiarism detection. *Intelligent Data Analysis*, 23(3):661–680.

Barg, A., Blakley, G. R., and Kabatiansky, G. A. (2003). Digital fingerprinting codes: Problem statements, constructions, identification of traitors. *IEEE Transactions on Information Theory*, 49(4):852–865.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2023). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

El-Rashidy, M. A., Mohamed, R. G., El-Fishawy, N. A., and Shouman, M. A. (2023). An effective text plagiarism detection system based on feature selection and svm techniques. *Multimedia Tools and Applications*, 83(1):2609–2646.

Esmaeili, M. M., Fatourechi, M., and Ward, R. K. (2011). A robust and fast video copy detection system using content-based fingerprinting. *IEEE Transactions on Information Forensics and Security*, 6(1):213–226.

Jiang, Y.-G., Jiang, Y., and Wang, J. (2014). Vcdb: A large-scale database for partial copy detection in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 357–371. Springer.

Jiang, Y.-G. and Wang, J. (2016). Partial copy detection in videos: A benchmark and an evaluation of popular methods. *IEEE Transactions on Big Data*, 2(1):32–42.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

King, D. (2007). Latest content id tool for youtube. Google Blog.

Kordopatis-Zilos, G., Tolias, G., Tzelepis, C., Kompatsiaris, I., Patras, I., and Papadopoulos, S. (2023). Self-supervised video similarity learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE.

Martemucci, M. and Swerdlow, A. (2017). Video game streaming brings new level of copyright issues. Law360. `https://www.law360.com/articles/920036/video-game-streaming-brings-new-level-of-copyright-issues`.

Nakatani, S. (2010). Language detection library for java.

Nie, X., Yin, Y., Sun, J., Liu, J., and Cui, C. (2017). Comprehensive feature-based robust video fingerprinting using tensor model. *IEEE Transactions on Multimedia*, 19(4):785–796.

Podilchuk, C. I. and Zeng, W. (1998). Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas in Communications*, 16(4):525–539.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Saeed, A. A. M. and Taqa, A. Y. (2022). Textual plagiarism detection using embedding models and siamese lstm. In *2022 International Conference for Natural and Applied Sciences (ICNAS)*, volume 31, page 95–100. IEEE.

Tomar, S. (2006). Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10.

Wary, A. and Neelima, A. (2018). A review on robust video copy detection. *International Journal of Multimedia Information Retrieval*, 8(2):61–78.

Yalcin, K., Cicekli, I., and Ercan, G. (2022). An external plagiarism detection system based on part-of-speech (pos) tag n-grams and word embedding. *Expert Systems with Applications*, 197:116677.

Zhang, D. Y., Li, Q., Tong, H., Badilla, J., Zhang, Y., and Wang, D. (2018a). Crowdsourcing-based copyright infringement detection in live video streams. In *Pro-*

ceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 367–374. IEEE.

Zhang, Y., Zhang, D., Vance, N., Li, Q., and Wang, D. (2018b). A lightweight and quality-aware online adaptive sampling approach for streaming social sensing in cloud computing. In *Proceedings of the IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 478–485. IEEE.