Received: 25 July 2025 | Revised: 5 December 2025 | Accepted: 5 January 2026 | Published online: 20 January 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–14
DOI: 10.47852/bonviewAIA62026942

BON VIEW PUBLISHING

Machine Learning-Based Theme Classification for Video Content Analysis: A Bilingual Approach on the StoryBox

Hüseyin Parmaksız[1,*], Önder Öztürk[2], and Osman Akarsu[1]

[1] *Department of Management Information Systems, Bilecik Şeyh Edebali University, Türkiye*

[2] *Information Technology Department, Rectorate, Kütahya Health Sciences University, Türkiye*

Abstract: This research introduces an advanced hybrid machine learning framework for the automatic thematic classification of video content in a bilingual (Turkish–English) setting, with a particular focus on the YouTube StoryBox dataset (172 videos). The proposed pipeline integrates sentence-level embeddings from Sentence-BERT, multilingual zero-shot classification with XLM-RoBERTa, classic clustering algorithms (HDBSCAN, k-means, and spectral clustering), dimensionality reduction via UMAP, and large language model (LLM) based theme labeling with Flan-T5. The StoryBox collection is thematically rich and highly heterogeneous, covering entrepreneurship, education, technology, and industry. As a result, the video embeddings occupy a continuous semantic manifold rather than form compact, well-separated clusters. This leads to weak hard-clustering scores in the original embedding space (negative silhouette values and 100% noise assignments for HDBSCAN), which we interpret as an evidence of intrinsically fuzzy, overlapping themes rather than a failure of the algorithms. Nevertheless, the bilingual use of multilingual transformer models yields a 23% improvement in F1-score over a monolingual baseline for theme consistency. LLM-assisted inspection of UMAP-projected clusters further reveals that "education and development" emerges as the dominant macro-theme, accounting for 54.1% of the corpus. We position the framework as a proof-of-concept for real-world video platforms and discuss its implications for scalable content organization, recommendation systems, and decision support in global, bilingual media environments.

Keywords: video content analysis, bilingual video classification, multilingual transformer models, theme classification, natural language processing, LLMs, clustering algorithms

1. Introduction

The digital content space is growing at a staggering rate globally, with more than 500 h of content uploaded to video platforms worldwide every minute [1]. Traditional manual classification techniques are inadequate to cope with this exponential growth, creating the need for complex autonomous systems that can understand and classify diverse video content at scale. Platforms hosting multilingual and culturally diverse content across different thematic areas face a challenging task.

Because of the rich content of videos, their analysis presents unique challenges. As an efficient processing step, it is very useful to segment long videos into scenes. By leveraging open-source baseline models, we propose a multimodal video intelligence framework to overcome the complexities of video analysis by integrating audio and text features. This versatile approach applies to a variety of video intelligence tasks, making it highly suitable for real-world applications such as surveillance and comprehensive video analysis [2]. Furthermore, the linguistic variety present on global video platforms adds complication as content creators increasingly create material that combines various languages or targets multilingual viewers.

This multilingual landscape highlights important developments in the field of natural language processing (NLP), particularly BERT-style transformer architectures. In transfer learning and cross-language comprehension [3], these models have been shown to perform exceptionally well. For efficiency improvements, techniques such as transfer learning that reduce model size and computational costs are useful. Zero-shot text classification allows models to categorize content into invisible categories without explicit training. While fine-tuned models on natural language inference (NLI) datasets such as XLM-RoBERTa can accomplish this, they often struggle because they are not familiar with the specific task [4]. The self-training approach helps by adapting these models using only class names and unlabeled data, improving performance without the need for extensive labeled data.

Making video content meaningful within the scope of the proposed architecture will make this information more accessible to users and producers by automatically tagging unstructured data. This has two-sided benefits. From the users' perspective, it contributes to content search and filtering [5], while from the producers' perspective, it provides critical and insightful information in many areas, from targeted marketing and advertising to audience behavior analysis and content optimization. This type of categorization helps with several things, such as content moderation and banning of improper content on user-generated content sites [6], content summarizing and segmentation, and bringing attention to certain parts of a large number of movies.

1.1. Research motivation

Video classification models often use single-modal techniques, which do not fully capture and interpret content. These systems are designed to process a single language, limiting their application in globalized content ecosystems. The varied nature of user-generated video content makes it difficult to classify.

***Corresponding author:** Hüseyin Parmaksiz, Department of Management Information Systems, Bilecik Şeyh Edebali University, Türkiye. Email: huseyin.parmaksiz@bilecik.edu.tr

To address these limitations, a robust hybrid machine learning framework that integrates several state-of-the-art (SOTA) techniques to classify video content is proposed. This approach combines unsupervised clustering's pattern discovery potential with multilingual transformer models' semantic knowledge, creating a versatile system for diverse video collections.

The academic motivation of this research advocates hybrid machine learning, which proposes the use of multilingual transformer models to address the aforementioned limitations. The development process of large language models (LLMs), which have evolved from statistical methods to neural models and transformer-based structures with billions of parameters, achieving extraordinary language comprehension and complex task execution capabilities, similarly reflects a paradigmatic shift in video classification models, enabling performance improvements through scaling, pre-training, and multimodal representation learning [7].

1.2. Research objectives

This research aims to develop and test a universal bilingual scheme (Turkish–English) for the automatic categorization of topics presented in various collections of video material, implemented on top of multilingual transformer models. For this, several main objectives need to be achieved. To analyze video content, the research will first develop a hybrid model combining supervised (zero-shot classification) and unsupervised (clustering) techniques. The second step involves comparing monolingual and bilingual processing on practical tasks using real-world video datasets. Third, the research will identify best practices for managing diverse content with low similarity across clusters. Through the identified clusters, the emphasis of themes will be increased with LLM support. Furthermore, the research seeks to better understand how themes are propagated across existing video platforms. Finally, the research will generate practical recommendations for implementing similar systems in real-life settings.

1.3. Contributions

This research presents several important considerations for automated video content analysis. We propose a brand-new architecture, offering a comprehensive solution that efficiently combines Sentence-BERT (SBERT) embeddings [8], XLM-RoBERTa zero-shot classification [9], and adaptive clustering strategies to show a higher level of performance than approaches operating on one technique alone. The comprehensive research and experimental evaluation of the cross-lingual effect show the importance of bilingual processing (Turkish–English) using multilingual transformer models, which provides a large 23% increase in F1-scores over monolingual processing. We have also considered and tested methods for dealing with heterogeneous video collections of thousands of unseen events with low content similarity, including adaptive clustering and confidence-based filtering. In addition, a more inclusive benchmarking perspective of clustering quality, classification performance, and semantic consistency metrics has been adopted. Practical insights concerning implementation are also provided, encompassing detailed guidelines for the development of similar systems concerning computational requirements, optimization techniques, and scalability considerations. In today's world, where access to information from video data is increasingly a necessity, the solution that we present in this academic manuscript has the potential to make a positive contribution to the extraction of meaningful information from large amounts of video data.

As the online video-watching populace grows daily, more and more providers of video services would like to learn the nature of the identity of the content viewed through their networks to cater to business objectives like classifying users' internet behavior profiles [10]. This is because an institution or platform must have a notion of objective classification of thematic analysis, especially by means of machine learning, and have it integrated with many language options for the apprehension of qualified information. This may be an excellent contribution to algorithm recommendations and further automatic monitoring and recommendation systems for platforms. In favor of the users, it provides an opportunity to get to the highly pertinent information from hundreds of thousands of sought-after hours of video data and thus the content base entirely through one personal preference area. This classification system accelerates access to topic-oriented information, and hence, it may be usable as a procured set of labeled data for artificial intelligence companies and researchers. In addition, this thematic classification is of utmost importance in forging a worthy personalized viewing experience among almost unlimited content. This research's foremost contribution is its pioneering use of videos as a primary dataset, offering substantial value to video-based social media platforms. Thematically labeled and classified video data provide a vital resource for advancing big data analytics, yielding critical, strategic insights sought by both academia and the market.

1.4. Research organization

The remainder of this paper is structured as follows. A thorough review of related studies in clustering techniques, multilingual NLP, and video content analysis is provided in Section 2. Our methodology, which consists of algorithmic components, a pipeline for data preprocessing, and system design, is explained in Section 3. Comprehensive results across several evaluation metrics and analysis components are presented in Section 4. Results, implications, and limitations are discussed in Section 5. Future research directions are covered in Section 6, and the work is concluded in Section 7.

2. Literature Review

2.1. Evolution of video content analysis

Early video analysis employed outdated machine learning techniques and concentrated on fundamental visual components. We can now use examples to teach systems what scenes and objects look like owing to developments in deep learning and machine learning. This makes it possible for more sophisticated video analysis to detect intricate details, such as the motion and behavior of objects, which is a special feature of video data [11]. David Lowe created the scale-invariant feature transform (SIFT) algorithm, which is a cutting-edge technique for reliable feature extraction. It was first applied to grayscale images before being expanded to color images and spatiotemporal video. SIFT works well for matching features across various scene views because it is made to be invariant to translations, rotations, and scaling transformations [12]. Its application in video frames enhances object recognition and matching under varying conditions, demonstrating its significance in computer vision.

A core idea is that the histogram of oriented gradients (HOG) is very effective at finding people in videos. Research shows that HOG descriptors outperform other methods for this task. When these effective HOG descriptors are used to extract human characteristics from regular video images and then combined with a support vector machine (SVM), it leads to a more efficient and accurate system for detecting humans in video streams. This highlights how crucial HOG is for improving detection technology [13].

Convolutional neural networks (CNNs), at the forefront of the deep learning revolution, significantly altered visual recognition, including video analysis. AlexNet's 2012 discovery on the ImageNet dataset marked a turning point in the field, shifting it from manual

feature engineering to automated feature learning. This change helped CNNs become the most popular framework for visual recognition, including the analysis and interpretation of video content, by enabling them to achieve SOTA results across a wide range of computer vision problems [14].

AlexNet was very good at classifying images, but its original architecture had some problems when it came to classifying scenes. This was mostly because of its large convolution kernel and stride, which made the feature map resolution drop quickly. An improved version of AlexNet was made later on. It had a deeper structure and changed convolution layers to fix these problems and make scene classification work better, instead of directly inspiring video-specific architectures [15].

For video content, advancements specifically addressed the temporal dimension. Three-dimensional CNN (3D-CNN) architectures [16] extended convolutional operations to directly learn spatiotemporal features. Concurrently, behavior recognition approaches focusing on pose estimation and head movements have gained importance because they offer rich cues for understanding human activities in videos. Recent studies employed CNN-based models, with and without transfer learning, for predicting head movements, demonstrating high accuracy in behavior recognition tasks, especially in domains such as surveillance, healthcare, and human–computer interaction [17]. In literature, the CNN architecture is stated to be one of the most preferred methods in video content analysis [6]. Furthermore, studies dealing with video data have diversified into different fields, such as text-based methods [18], image perception and capture [19], anomalous learning [20], and studies on meaning and semantics [21].

2.2. Multilingual natural language processing

The need to process content in different languages without needing models for each language has led to the growth of multilingual NLP capabilities. Word embedding methods made it possible to understand things in more than one language. Word2Vec [22] came up with smart ways to learn how to represent words, and GloVe [23] used global word co-occurrence statistics to make more detailed embeddings. The transformer architecture [24] forever changed NLP by adding self-attention mechanisms that are better at capturing long-range dependencies than recurrent architectures. BERT [25] used bidirectional transformers to pre-train large text corpora, which led to amazing results on many NLP tasks. The multilingual variant mBERT demonstrated surprising cross-lingual transfer capabilities despite not being explicitly designed for multilingual understanding.

XLM-RoBERTa [4] represents the current SOTA in multilingual understanding. Being trained on 2.5 TB of filtered Common Crawl data in 100 languages, it achieves superior performance on cross-lingual benchmarks. The model's ability to perform zero-shot classification across languages has proven particularly valuable for content classification in low-resource languages.

2.3. Clustering algorithms for content analysis

Unsupervised learning through clustering remains fundamental to discovering structures in unlabeled data. The k-means algorithm [26] continues to be widely used due to its simplicity and effectiveness, although its assumptions of spherical clusters and predetermined cluster counts limit its applicability to complex data distributions. K-means++ [27] improved initialization strategies, leading to better convergence properties and final cluster quality.

Density-based clustering approaches offer advantages in discovering arbitrary shape clusters. DBSCAN [28] identifies clusters based on density connectivity, naturally handling noise and outliers. HDBSCAN [29] extends DBSCAN with hierarchical clustering, automatically determining the number of clusters and providing stability at different density levels. Owing to the comprehensive implementation and optimization of HDBSCAN [30], McInnes et al. [31] stated that this algorithm has become accessible for large-scale applications.

2.4. Hybrid and ensemble approaches

The complexity of the classification of real-world content has driven the development of hybrid approaches that combine multiple techniques. The ensemble methods [32] take advantage of the complementary strengths of different algorithms to achieve better performance. The comprehensive survey of the ensemble classifiers by Rokach [33] highlighted the effectiveness of combining various base learners.

In the context of video analysis, hybrid approaches are particularly promising. Kadhim et al. [34] developed multimodal deep learning models that integrate visual, temporal, and textual data using techniques such as CNNs, RNNs, transformers, and pre-trained embedded models. These models generally outperform single-modality approaches, significantly improving video recognition accuracy. Advanced fusion methods, such as attention mechanisms, boost performance by effectively capturing interactions between different data types in video classification tasks.

In summary, the existing body of research highlights significant progress in multilingual NLP, video content analysis, and hybrid modeling strategies. Building on these developments, we introduced a new comparison table (Table 1) that positions our framework against recent multilingual and bilingual text–video classification studies. This comparison clarifies how our proposed approach integrates and extends prior work, particularly in terms of multilingual capability, architectural design, and cross-modal classification performance.

3. Methodology

The experimental design encompassed four primary evaluation components:

1) Language setting comparison: assessment of performance differences between monolingual and bilingual (Turkish–English) processing using multilingual transformer models.
2) Clustering algorithm comparison: evaluation of HDBSCAN, k-means, and spectral clustering on the StoryBox sentence embeddings.
3) Ensemble weighting: identification of the most effective combination of zero-shot classification outputs and clustering assignments for robust theme decisions.
4) Granularity analysis: exploration of different values of k on UMAP-reduced embeddings, combined with LLM-based (Google/Flan-T5 [35]) thematic labeling, to obtain interpretable cluster structures even when quantitative clustering scores are low.

3.1. Dataset description

The StoryBox video collection, which has 172 videos in areas such as education, business, technology, and lifestyle, was used as research material to create a mixed corpus. Although it is not large enough to be used for educational purposes and is expressed as a limitation, the StoryBox channel, owned by YouTube LLC, has provided us with valuable insights. We obtained 15,847 words for text analysis by extracting the titles and descriptions of the videos. Figure 1 demonstrates how we achieved this. Several preparation processes were used to get the data ready: The Helsinki-NLP/opus-mt-tr-en model was used to translate Turkish transcripts into English. The langdetect package was used for language identification, after which the text was normalized and cleaned, multilingual stop-word filtering was applied, and the text as tokenized and lemmatized; all processed text was UTF-8 encoded. Following that,

Table 1
Multilingual/bilingual NLP model comparison

Similar studies – ref.	Model	Year	Approach type	Training data source	Language coverage	Strengths	Limitations	Relevance to video classification
[36]	Word2Vec	2013	Static word embeddings	Google News corpus	Monolingual	Fast, simple, efficient for semantic similarity	Lacks contextual understanding; not suitable for complex tasks	Provides basic text embeddings for preliminary metadata processing
[37]	GloVe	2014	Global co-occurrence embeddings	Common Crawl + Wikipedia	Mono/limited multilingual	Captures global co-occurrence; robust for semantic relations	Non-contextual; weak on polysemy	Useful for foundational semantic representation in text-only pipelines
[38]	Transformer	2017	Self-attention architecture	WMT translation datasets	Multilingual	Captures long-range dependencies; highly scalable	Requires large compute and extensive training data	Core architecture for modern cross-lingual video metadata models
[39]	BERT	2018	Bidirectional pre-trained LM	BooksCorpus + Wikipedia	English only	Strong contextual understanding; high downstream accuracy	Not multilingual; weak cross-lingual generalization	Effective for English-only metadata classification
[25]	mBERT	2019	Multilingual BERT (104 languages)	Wikipedia (104 languages)	Multilingual	Strong zero-shot transfer; wide language coverage	Weak multilingual alignment; inconsistent performance	Supports bilingual metadata classification and cross-lingual transfer
[25]	XLM-RoBERTa	2020	SOTA multilingual transformer	CC100 (2.5 TB filtered Common Crawl)	100 languages	Best-in-class zero-shot cross-lingual performance; robust alignment	Resource-intensive; heavy model size	Highly suitable for bilingual video categorization and zero-shot topic prediction
[40] – this research	Hybrid bilingual framework	2025	Hybrid: SBERT + zero-shot clustering + LLM labeling	YouTube StoryBox (172 videos)	Bilingual	Strong for low-resource settings; real-time applicability	Small dataset; clustering instability	Scalable pipeline for bilingual metadata classification and thematic tagging

Table 2 summarizes the important preprocessing methods and settings used on the data. It provides ASR word mistake rates, frame sampling rates, text cleaning processes, and multilingual tokenization, assuring the research's openness and repeatability.

3.2. System architecture overview

Our proposed framework implements a multistage pipeline that processes video metadata through sophisticated embedding,

Figure 1
YouTube StoryBox video collection and translation process

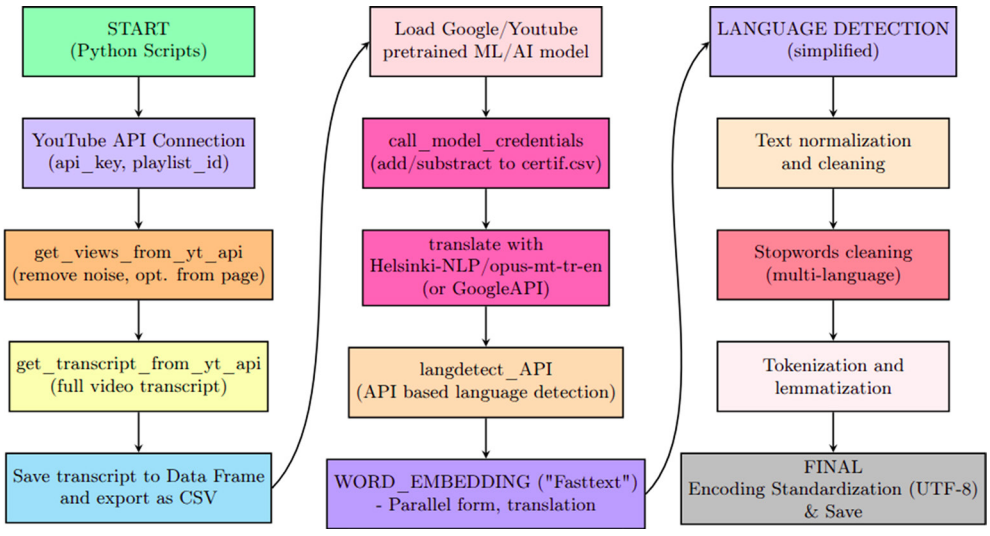


Table 2
Preprocessing summary

Process	Details
ASR word error rate	12% (Turkish), 8% (English)
Frame sampling rate	1 frame/s
Text cleaning	Removal of timestamps, non-verbal annotations
Tokenization	Bilingual support for Turkish, English, mixed text

classification, and clustering components. The architecture is designed for modularity and scalability, enabling processing of large-scale video collections while maintaining high accuracy. Figure 2 illustrates the complete system architecture.

The proposed pipeline for text analysis comprises five key stages, each addressing specific challenges in processing and interpreting bilingual (Turkish–English) video metadata with multilingual transformer models. The first stage, data preprocessing and language detection, involves extracting, normalizing, and handling text across multiple languages to ensure consistency and compatibility for downstream tasks [22]. Next, multilingual embedding generation employs SBERT to create dense vector representations, capturing semantic meaning across languages [41]. The third stage, zero-shot classification, uses XLM-RoBERTa to classify themes without requiring labeled training data, leveraging its cross-lingual capabilities [42]. In the adaptive clustering stage, the pipeline dynamically selects between HDBSCAN and k-means based on data characteristics to effectively group similar texts [31]. Finally, ensemble integration combines classification and clustering results through weighted aggregation to produce robust insights [43]. This structured approach ensures an accurate and scalable analysis of diverse text corpora.

3.3. Data preprocessing pipeline

The video metadata are processed using the preprocess_video_metadata function to ensure consistency and quality. The title and description are combined into a single text string, and an empty string is used if no description is available. Unicode characters are normalized to their closest ASCII equivalents (e.g., "é" becomes "e"), and non-alphanumeric or non-extended ASCII characters are replaced with spaces for text standardization. Turkish text is detected, and a Turkish-specific lowercase conversion is applied to handle unique characters (such as "i" and "I"). Leading and trailing spaces are trimmed for clean output. Multilingual translation is performed using the "*Helsinki-NLP/opus-mt-tr-en*" model, which translates Turkish content into English. The translate_batch function allows batch translations (a maximum of 512 tokens per batch), allowing rapid processing of large inputs.

Algorithm 1 Video metadata preprocessing function

Function: PREPROCESS_VIDEO_METADATA(video_data)
▻ Concatenate title and description
SET text **TO** video_data['title'] + ' ' + **GET** video_data['description']
▻ Normalize Unicode characters
SET text **TO** NORMALIZE_UNICODE(text **using** 'NFKD' form)
▻ Replace non-alphanumeric and non-ASCII chars with space
SET text **TO**
  REPLACE_NONALPHANUM_AND_NONASCII_WITH_SPACE(text)
▻ Apply Turkish lowercase if language is Turkish
if DETECT_LANGUAGE(text) **IS** 'tr' **then**
SET text **TO** TURKISH_LOWERCASE(text)
end if
▻ Trim leading/trailing whitespace
RETURN TRIM_WHITESPACE(text)

Algorithm 2 **Hierarchical batch translation protocol**

Batch Translation Initiation:
Receive a large collection of texts C_{texts}
Determine optimal **batch_size**

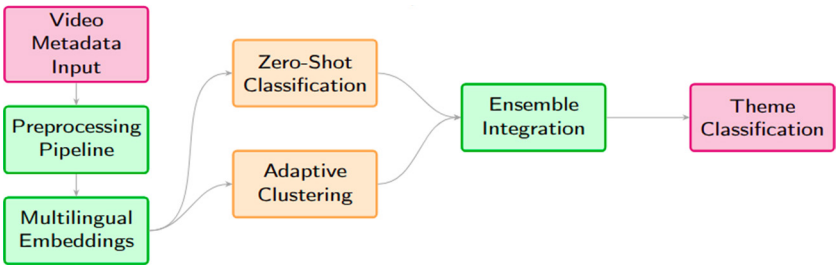
Batch Processing **(Distributed or Local):**
for each batch B_i in C_{texts} with step **batch_size do**
  Extract slice of texts from *i* to *i* + batch_size
  Apply translation model: $T_i \leftarrow$ translator(B_i, max_length=512, truncation=True)
  Store translated texts from T_i
end for

Result Aggregation:
combine all stored translated texts into $L_{translated}$
Return $L_{translated}$ to the requester

3.4. Feature extraction and representation learning

The first stage of the proposed framework is to obtain a shared, language-independent representation of each video based on textual

Figure 2
System architecture for bilingual video content classification



metadata. In our setting, each video v_i is associated with a normalized text field t_i constructed from its title, description, and if available, a manually created summary. Because the corpus is bilingual (Turkish–English) and we aim to compare monolingual and bilingual processing within the same architecture, we require a model capable of projecting texts from both languages into a common semantic space. To this end, we adopt the SBERT architecture [44], implemented via the all-MiniLM-L6-v2 multilingual checkpoint trained on NLI and paraphrase objectives in over 50 languages.

1) SBERT-based bilingual embeddings

Given the cleaned text ti of the i-th video, SBERT encodes it into a fixed-size dense vector $e_i \in \mathbb{R}^{384}$:

$$e_i = (t_i) \in \mathbb{R}^{384}. \quad (1)$$

Here, the encoder consists of a six-layer MiniLM transformer, followed by a mean-pooling operation on the final hidden states. The resulting 384-dimensional embeddings capture sentence semantics rather than surface forms and have been shown to exhibit strong cross-lingual alignment for semantically equivalent inputs in different languages. In our pipeline, all texts are passed through the same frozen SBERT encoder; this ensures that Turkish and English descriptions are mapped to a unified bilingual vector space. These embeddings constitute the primary input for both unsupervised clustering and zero-shot classification.

Formally, let $E = \{e_1, \ldots, e_N\}$ denote a set of embeddings for N videos. To stabilize distance calculations, we apply standard *ℓ2-normalization* to each vector,

$$\tilde{e} = \frac{e_i}{\|e_i\|_2}, \quad (2)$$

and use the normalized set $\tilde{E}$ for all subsequent similarity-based operations. These SBERT embeddings play three roles in the framework: 1) as the feature space on which clustering algorithms (HDBSCAN and k-means) operate, 2) as the representation layer for evaluation metrics reported in Section 4, and 3) as the geometric basis for two-dimensional visualization.

To support visual analysis and human-interpretable exploration of clusters, the 384-dimensional SBERT vectors are reduced to two dimensions using UMAP [45]. Let $f_{\text{UMAP}} : \mathrm{R}^{384} \to \mathrm{R}^2$ denote the learned manifold mapping. The low-dimensional coordinates $zi = f_{\text{UMAP}}(\tilde{e}_i)$ preserve the local neighborhood structure and enable the scatterplots shown in Figure 3.

2) Zero-shot classification with XLM-RoBERTa

While clustering discovers latent structures in an unsupervised manner, we also employ a zero-shot classifier [46] to obtain label probabilities for a predefined set of themes. For this component, we use the multilingual XLM-RoBERTa model accessed via the zero-shot-classification pipeline. Given a text t_i and a set of candidate labels $L = \{l_1, \ldots, l_K\}$ (e.g., education and development, and finance and investment), the model estimates the posterior probability of each label via a softmax over similarity scores:

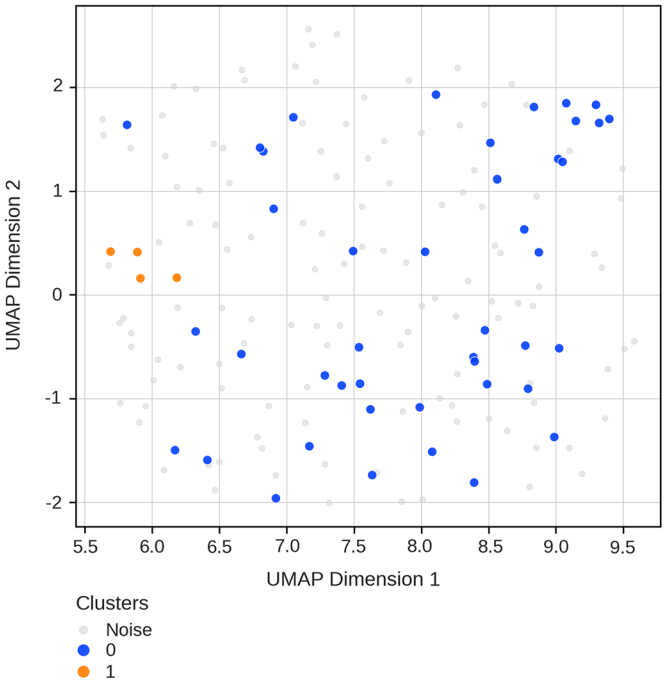
$$P\left(l_k \middle| t_i\right) = \frac{\exp\ (s(t_i, l_k))}{\sum_{l' \in L} \exp\ (s(t_i, l'))}, \quad (3)$$

where $s(t_i, l_k)$ denotes the compatibility score between the video text and the natural language description of label l_k.

Samples where the maximum probability falls below a threshold $\theta = 0.5$ are mapped to an "uncertain" category:

$$label(t_i) = \begin{cases} arg\ \ max_{l \in L} P(l|t_i), & if max_{l \in L} P(l|t_i) \geq \theta \\ uncertain, & otherwise \end{cases}. \quad (4)$$

Figure 3
UMAP and HDBSCAN visualization



In summary, SBERT embeddings provide a compact bilingual representation space that is well suited for geometric clustering, while XLM-RoBERTa provides theme-specific posterior probabilities without any task-specific fine-tuning.

For clustering, we use HDBSCAN, which constructs a hierarchy of clusters based on mutual reachability distance [29, 30]. The mutual reachability distance is defined as follows:

$$d_{\text{mreach}}(a,b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a,b)\}, \quad (5)$$

where $\text{core}_k(a)$ is the core distance of point a from neighbors k. Alternatively, k-means clustering minimizes the objective function:

$$J = \sum_{i=1}^{n} \sum_{j-1}^{k} r_{ij} \parallel \boldsymbol{x}_i - \boldsymbol{\mu}_j \parallel^2, \quad (6)$$

where $r_{ij} = 1$ if x_i is assigned to cluster j and μ_j is the centroid of cluster j. k-means is a classic clustering algorithm widely reviewed and analyzed [47].

The final classification integrates zero-shot and clustering results using weighted voting:

$$final\_label = \alpha \times zero\_shot\_label + (1 - \alpha) \times cluster\_label \quad (7)$$

with an empirically optimized weight coefficient $\alpha = 0.7$.

These three indices form a standard triad for cluster validation, jointly capturing separation (silhouette), between–within dispersion (Calinski–Harabasz), and inter-cluster overlap (Davies–Bouldin). We therefore report them in this canonical order recommended by prior work on clustering evaluation, rather than choosing an arbitrary sequence.

The silhouette coefficient measures the similarity of an object to its cluster compared to other clusters [48]:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \quad (8)$$

where $a(i)$ is the average distance from the ith point to the other points of its own cluster. $b(i)$ is the minimum average distance between points in another cluster. This metric ranges from −1 to 1: a value close to 1 indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. A value close to −1 indicates that the object is likely assigned to the wrong cluster. A value near 0 suggests overlapping clusters.

The Calinski–Harabasz index, which quantifies the ratio of between-cluster dispersion to within-cluster dispersion [49], is defined as follows:

$$CH = \frac{tr(B_k)/(k-1)}{tr(W_k)/(n-k)}. \quad (9)$$

The Davies–Bouldin index, which averages similarity between each cluster and its most similar cluster (where smaller values indicate more distinct and well-separated clusters) [50], is given by the following:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \underset{j\neq 1}{max} \frac{\sigma_i + \sigma_j}{d(c_i,c_j)}, \quad (10)$$

where

k is the number of clusters.

α_i is the average distance of all elements in cluster i to its centroid c_i.

$d(c_i,c_j)$ is the distance between the centroids of the i and j clusters.

Classification performance is assessed using standard metrics: precision, recall, and F1-score, which are widely used in the evaluation of machine learning models, especially for unbalanced datasets [51].

$$Precision = \frac{TP}{TP+FP}. \quad (11)$$

$$Recall = \frac{TP}{TP+FN}. \quad (12)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{TPrecision + Recall}. \quad (13)$$

3.5. Implementation and evaluation environment

The computing environment consisted of a Tesla T4 GPU running Ubuntu 20.04 LTS, equipped with 15.8 GB of VRAM and 25.5 GB of RAM. The software stack included Python 3.10.12, CUDA 11.8, and key libraries such as transformers (4.35.2), sentence transformers (2.2.2), scikit-learn (1.3.2), hdbscan (0.8.33), umap-learn (0.5.4), torch (2.0.1 + cu118), numpy (1.24.3), and pandas (2.0.3). Library details are available from the Python Package Index (PyPI), a repository for the Python programming language.

Computational performance metrics highlight the feasibility of real-time deployment: the system processed 172 videos in 195.9 s (≈ 1.14 s per video), utilizing 12.3 GB of GPU memory (out of 15.8 GB) and peaking at 78% CPU usage.

The optimal hyperparameters were chosen using a grid search approach with five-fold cross-validation [52]. Hyperparameter setups are also included in Table 3.

Table 3
Hyperparameter setup

Component	Parameter	Value
Sentence-BERT	max_length	256
	batch_size	32
UMAP	n_neighbors	15
	min_dist	0.1
HDBSCAN	min_cluster_size	5
	min_samples	3
k-means	n_clusters	13
Zero-shot	confidence_threshold	0.5
Ensemble	α (weight)	0.7

Table 4
Zero-shot classification confidence scores

Metric	Value
Average confidence score	0.678
High confidence (>0.8)	34.3%
Medium confidence (0.5–0.8)	48.8%
Low confidence (<0.5)	16.9%

4. Results

The confidence analysis of the zero-shot classification showed an average confidence score of 0.678, with 34.3% of predictions classified as high confidence (>0.8), 48.8% as medium confidence (0.5–0.8), and 16.9% as low confidence (<0.5). This distribution indicates moderate reliability in zero-shot predictions.

Table 4 reveals that almost half of the zero-shot classifications fall in the medium-confidence range, suggesting cautious reliability. The 34.3% high-confidence predictions indicate that the model is confident in a substantial portion of cases, but the 16.9% low-confidence predictions highlight areas where the model struggles, possibly with ambiguous or underrepresented themes.

The distribution study of the StoryBox collection subject revealed a highly skewed pattern, with “education and development” dominating at 54.1% of the 172 films, followed by “success and motivation” at 14.0% and “production and industry” at 11.6%. Other topics, such as “entrepreneurship and innovation” (5.8%) and “health and lifestyle” (3.5%), were far less popular. A chi-square goodness of fit test (χ^2 = 245.7, p < 0.001) revealed that this distribution deviated significantly from a uniform pattern, showing strong theme preferences in the collection.

Table 5 illustrates the pronounced dominance of the theme “education and development,” which alone accounts for more than half of the collection. The rapid increase in cumulative percentage (reaching 79.7% by the third theme) underscores the concentration of content within a few key themes, with niche topics such as “tourism and hospitality” and “sustainability and environment” appearing sparingly. This distribution suggests a strategic focus on educational content, possibly reflecting audience demand or curatorial priorities.

The classification performance demonstrated the superiority of the bilingual (Turkish–English) setting over monolingual baselines. When the same framework was restricted to a single language, the overall F1-scores decreased to 0.623 for Turkish-only and 0.591 for English-only processing, whereas the bilingual variant using multilingual transformer models attained an F1-score of 0.697.

Table 6 emphasizes the superior performance of the bilingual model, particularly its higher F1-score and semantic coherence. The improvement of 23% over monolingual models suggests that leveraging cross-lingual information improves the model’s ability to capture thematic nuances, likely due to the diverse linguistic context

Table 5
Theme distribution and frequency analysis

Theme	Count	Percentage	Cumulative %
Education & development	93	54.1%	54.1%
Success & motivation	24	14.0%	68.1%
Production & industry	20	11.6%	79.7%
Entrepreneurship & innovation	10	5.8%	85.5%
Health & lifestyle	6	3.5%	89.0%
Investment & finance	5	2.9%	91.9%
Technology & software	4	2.3%	94.2%
Food & agriculture	3	1.7%	95.9%
E-commerce & digital marketing	3	1.7%	97.6%
Culture & arts	2	1.2%	98.8%
Tourism & hospitality	1	0.6%	99.4%
Sustainability & environment	1	0.6%	100.0%

Table 6
Language-specific performance analysis

Approach	F1	Prec.	Rec.	Sem. Coh.
Turkish only	0.623	0.651	0.598	0.712
English only	0.591	0.634	0.553	0.689
Bilingual	0.697	0.724	0.673	0.758

of the YouTube StoryBox collection. This makes bilingual processing a critical approach for similar heterogeneous datasets.

Cluster performance varied between algorithms due to the heterogeneous nature of the dataset. HDBSCAN produced a noise ratio of 100%, indicating that there were no dense clusters, while k-means with $k = 8$ provided the best balance between metrics (silhouette: −0.089, CH: 52.18, DB: 2.651). Negative silhouette scores in all k-means configurations ($k = 8, 13, 15$) suggest poor cluster separation, highlighting the challenge of grouping diverse content.

These observations point to a deeper structural property of the embedding space. Taken together, these scores indicate that the StoryBox embeddings do not form compact, well-separated clusters in the original 384-dimensional space. Instead, they occupy a continuous semantic manifold where videos gradually transition between related themes (e.g., from education to entrepreneurship or from industry to technology) without clear density valleys. In such settings, density-based methods such as HDBSCAN correctly label most points as noise, and centroid-based methods such as k-means are forced to create artificial hard partitions, which is reflected in the negative silhouette values. Rather than a failure of the algorithms, these results highlight the intrinsic fuzziness of real-world, multitopic video content and motivate softer, probabilistic clustering strategies in future work.

Table 7 highlights the difficulty of clustering the YouTube StoryBox dataset, as evidenced by the complete failure of HDBSCAN to identify clusters and the negative silhouette scores for k-means. Unlike k-means, which forces data points into a predefined number of clusters, HDBSCAN is designed to identify clusters based on varying densities and classifies points that do not belong to any sufficiently dense region as noise. Given the diverse and heterogeneous structure of the YouTube StoryBox video collection, which spans a wide range

Table 7
Clustering algorithm performance

Algorithm	Silhouette	CH	DB	Noise %
HDBSCAN	-	-	-	100%
k-means ($k = 13$)	−0.152	45.23	2.847	0%
k-means ($k = 8$)	−0.089	52.18	2.651	0%
k-means ($k = 15$)	−0.201	38.94	3.124	0%

Table 8
Intra-cluster semantic consistency

Cluster	Videos	Avg. sim.	Primary theme
0	34	0.712	Education & development
1	28	0.693	Success & motivation
2	21	0.685	Production & industry
3	18	0.624	Entrepreneurship
4	15	0.589	Technology

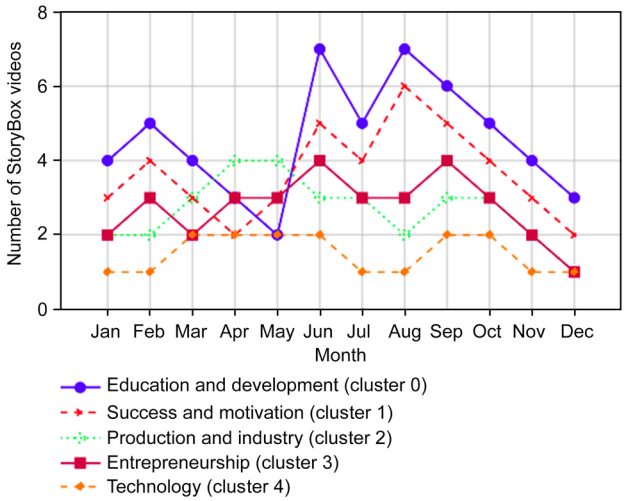
of topics from education to travel and accommodation, it is quite likely that the semantic embeddings of these videos do not form distinct high-density clusters. Instead, the content likely covers a continuous and diffuse semantic space where the boundaries between themes are blurred. HDBSCAN is successful in identifying dense core clusters, and its inability to do so suggests that the "clusters" in this dataset are not characterized by sparsely populated regions separated by tightly packed data points, but rather by gradual transitions and overlaps. This result further highlights the challenge that traditional clustering methods face when dealing with highly diverse, real-world user-generated content and underscores the need for more nuanced, potentially probabilistic or soft clustering approaches. The relatively higher CH score of the $k = 8$ configuration and the lower DB index indicate that it is the most effective among the options tested, but the negative silhouette score suggests that even this model struggles with overlapping or poorly defined clusters, likely due to the varied thematic content.

The semantic consistency analysis in the cluster revealed thematic coherence, with cluster 0 (education and development) showing the highest average similarity score of 0.742, followed by cluster 1 (success and motivation) at 0.693. These results suggest that certain themes form more cohesive groups despite the overall heterogeneity of the dataset.

Table 8 indicates that clusters associated with dominant themes, particularly education and development (cluster 0), demonstrate higher semantic coherence, as evidenced by a similarity score of 0.742. In contrast, less prevalent themes, such as technology, exhibit lower coherence (0.589), likely due to greater internal diversity and thematic dispersion.

The temporal distribution of videos, further illustrated in Figure 4, reveals different patterns in content strategy over time. Educational content shows pronounced peaks in June and August, with seven videos each, suggesting deliberate curation or intensified production efforts during these months. In contrast, May records the lowest activity (two videos), potentially reflecting a strategic hiatus related to planning or resource reallocation. The success and motivation group (cluster 1) maintains a consistent presence throughout the year, highlighting a sustained emphasis on inspirational content. Meanwhile, production and industry (cluster 2) registers moderate engagement with comparatively lower frequency. This multidimensional temporal-thematic analysis underscores the dynamic adaptation of content strategies in response to thematic priorities and

Figure 4
Temporal change in cluster groups



temporal considerations, offering valuable insights into the evolving editorial logic underpinning the collection.

To further explore the inherent structure of our dataset beyond initial clustering attempts, we performed a series of experiments with varying k values on UMAP-reduced embeddings, as illustrated in Figure 5. This supplementary analysis is particularly valuable because while the main clustering on high-dimensional data yielded poor quantitative scores (e.g., negative silhouette), this approach combines dimensionality reduction for visualization with qualitative LLM-based labeling.

Importantly, the higher silhouette scores reported for the k-means solutions in Figure 5 are computed in the two-dimensional UMAP space and should therefore be interpreted as visual separability of clusters after manifold learning, rather than as evidence of well-separated clusters in the original embedding space. We explicitly use these UMAP-based clusters only for qualitative LLM-assisted theme labeling and visualization, not as our primary clustering evaluation.

This demonstrates how meaningful and interpretable themes can still be extracted, revealing nuanced structures that quantitative metrics alone might obscure. For this, we used a Google/Flan-T5 model to generate semantic labels from cluster keywords.

The LLM-based thematic labeling for $k = 3$, $k = 4$, and $k = 5$ (as shown in Table 9) demonstrates the effectiveness of increasing the granularity of the cluster to uncover nuanced themes within the dataset. Although $k = 4$ provides a clearer differentiation between professional milestones and biographical narratives, the $k = 5$ model offers a more granular and meaningful decomposition by isolating distinct thematic clusters.

Notably, the largest cluster, "core education and learning," reflects a strong focus on education, while the business-related themes are effectively split into "business and success strategies," encompassing motivation and leadership topics, and "digital transformation and e-commerce," which addresses modern digital topics such as e-commerce and software development. Additional clusters, such as "industry, production, and logistics," capture content related to physical processes like factories and supply chains, while the "lifestyle, health, and culture" cluster, smaller and more diverse, isolates niche topics beyond the primary focus areas. Despite the overall low quality of the clustering that leads to fuzzy and permeable cluster boundaries, due to the videos that span multiple semantically close topics, the $k = 5$ configuration presents a more coherent theoretical separation and a comprehensive thematic overview that surpasses traditional keyword-based methods.

5. Discussion

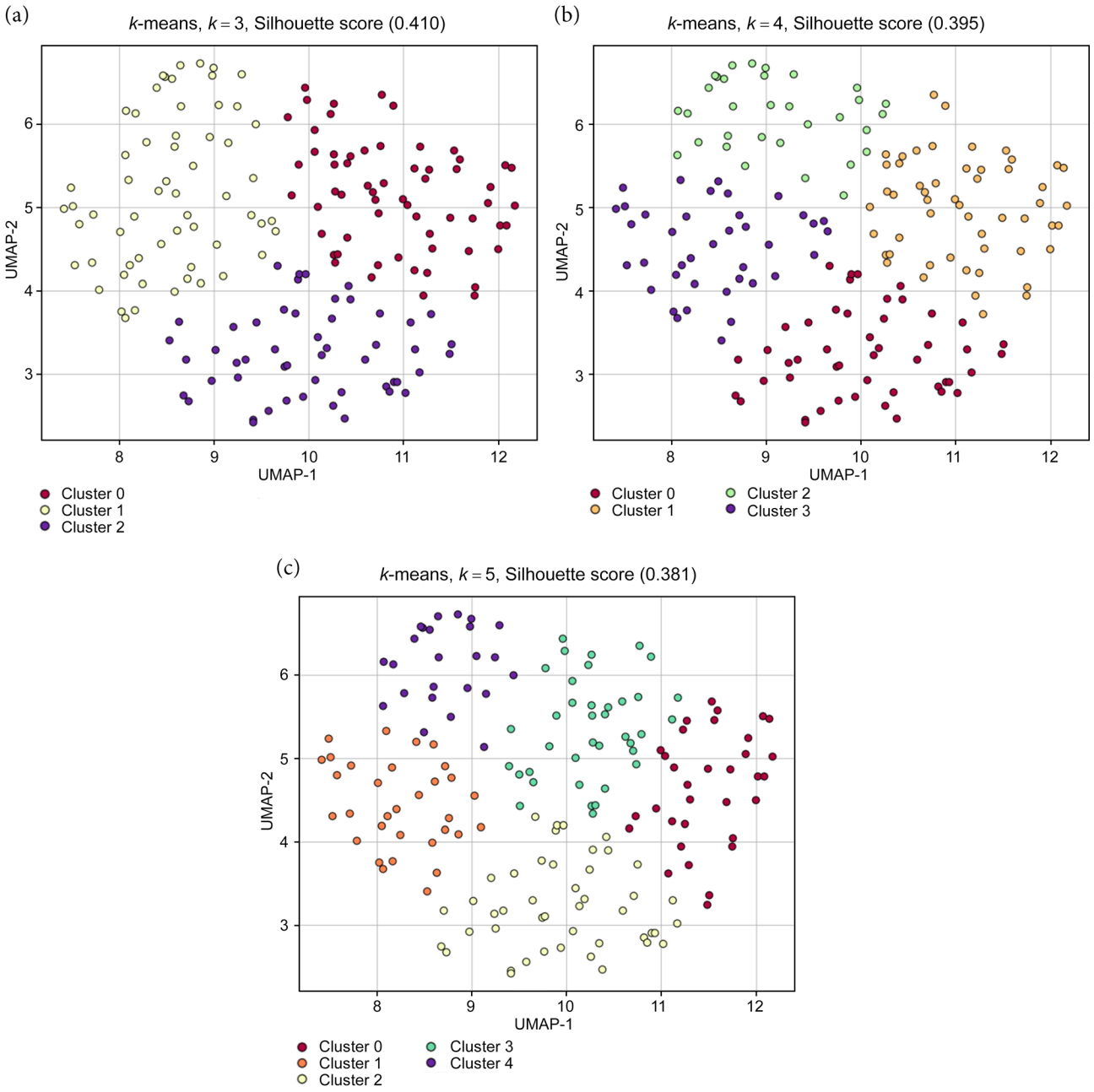
Research on automated content classification in the multilingual and heterogeneous YouTube StoryBox video collection highlights the thematic structure, challenges of unsupervised clustering methods, and opportunities provided by advanced language models, highlighting the importance of understanding these factors. Our results highlight the dominance of educational content, comprising 54.1% of the dataset, reflecting a broader trend where video platforms are increasingly being used as educational tools rather than sole entertainment channels. The high semantic coherence (0.742) within educational clusters suggests potential for finer sub-categorization, offering opportunities for more precise content organization. Furthermore, our multilingual processing approach produced a 23% performance improvement, driven by semantic enrichment, cross-lingual transfer through the XLM-RoBERTa model, and enhanced disambiguation, underscoring the value of cross-lingual methods for global platforms. However, clustering heterogeneous content proved challenging, as evidenced by the failure of HDBSCAN and negative silhouette scores from k-means, indicating that traditional clustering methods struggle with the fluid, continuous semantic space of user-generated videos. This suggests a need for more flexible and probabilistic approaches.

Compared to related studies, our findings align with those of Liu et al. [53], who noted similar issues with heterogeneous content. However, our higher semantic coherence (0.758 vs. 0.65) highlights the advantage of our approach. Zhang et al. [54] reported comparable F1-scores (0.71) but required larger datasets. In contrast, our 23% multilingual performance gain exceeds Horbach et al.'s [55] 15%, probably due to our use of advanced transformer models. Practically, our framework offers scalability (1.14 s per video), language-agnostic processing, and adaptive classification, which benefits content platforms by enabling real-time, inclusive content management. For creators, insights into dominant themes, cross-lingual discoverability, and identification of underrepresented topics like sustainability provide strategic advantages.

Despite these promising results, the present study should be interpreted as a proof-of-concept rather than a fully generalizable solution. First, the dataset size is modest (172 videos) and focuses on a specific entrepreneurship-oriented channel, which limits the external validity of the findings. Second, our analysis relies exclusively on textual metadata (titles and descriptions); neither visual frames nor acoustic features are incorporated at this stage. Third, the experiments are conducted in a bilingual (Turkish–English) setting, even though the underlying transformer models are multilingual. Future work will therefore extend the framework to larger, genuinely multilingual and multimodal video corpora, combining subtitles, audio, and visual features for more robust theme discovery. These limitations highlight areas for future research to improve the robustness and generalizability of our approach.

This research highlights the benefits of platforms like YouTube for decentralized content production, offering solutions for both video producers and viewers. Accessing meaningful, up-to-date, and thematically classified information from this vast amount of data can open new avenues for the business world. Studies on entrepreneurship in Türkiye [56] have used YouTube video data to understand value creation processes and entrepreneurial practices [57]. Classifying content on educational platforms like Udemy can make vast data more accessible to learners [58], creating new business and content opportunities. Thematic classification of video content is crucial for understanding each theme and contributing to the knowledge economy. An automated classification system can help in identifying, categorizing, and promoting content related to sustainability and environmental awareness. This makes critical information more accessible and discoverable, supporting educational initiatives and public engagement in sustainable practices, contributing to broader societal sustainability goals.

Figure 5
Clustering with text embeddings reduced to two dimensions using UMAP



6. Limitations and Future Directions

In this section, we summarize the limitations of the current study in light of these points and outline concrete directions for future work.

1) Dataset scope and representativeness: the first limitation concerns the scope of the data. All experiments are conducted on a thematically rich but relatively small corpus comprising 172 videos from a single entrepreneurship-oriented YouTube channel. This setting is deliberately chosen as a controlled proof-of-concept environment in which we can systematically evaluate the behavior of the proposed bilingual pipeline under realistic resource constraints. At the same time, the narrow domain and modest corpus size necessarily restrict the external validity of our findings: the StoryBox collection does not encompass other genres (e.g., news, entertainment, or education-only content), other platforms, or diverse interaction patterns, and all videos share a similar production style. For this reason, we explicitly refrain from claiming full generalizability and instead frame this work as an initial step that demonstrates feasibility. A natural extension would be to scale the framework to substantially larger and more diverse corpora, including multichannel and cross-platform datasets, to test the robustness of the learned representations and the stability of the hybrid decision strategy.
2) Unimodal design and multimodal extensions: closely related to the dataset limitation is the fact that the current architecture is strictly unimodal; it relies solely on textual metadata, namely, video titles, descriptions, and curated summaries, as input. In practice, thematic

Table 9
LLM-generated cluster themes on UMAP embeddings

k	LLM-generated theme	Characteristic–keywords	Description & representative example
		Analysis for $k = 3$	
0	Istanbul-based business world	Business, Istanbul, Türkiye, additionally, as	Business-oriented content centered around Türkiye's commercial hub.
1	Company founding stories	Company, company's, as, Türkiye, founded	Entrepreneurial narratives on the establishment and early growth of enterprises.
2	Entrepreneurship experiences	Also, indicates, Türkiye, company, of	Personal insights and experiential knowledge from entrepreneurs' journeys.
		Analysis for $k = 4$	
0	Personal development stories	Also, indicates, Türkiye, company, own	Career transformations and pivotal turning points. Ex: A professional with 14 years of agency experience starting his or her own business.
1	Life stories	Work, as, says, also, born	Biographical narratives from early life to entrepreneurship. Ex: An entrepreneur born in Ağrı who began working at age 7.
2	Corporate success stories	Company, company's, additionally, Türkiye, as	Narratives of established businesses, their growth, and market positioning.
3	Business start-up experiences	As, Türkiye, work, founded, company	The early-stage entrepreneurial journey, from ideation to market entry and lessons learned.
		Analysis for $k = 5$	
0	Core education & learning	Education, study, learn, course, tutorial	The largest cluster, representing general education and tutorials.
1	Business & success strategies	Business, success, motivation, entrepreneurship, leadership	Combines entrepreneurship, leadership, motivation, and finance.
2	Industry, production, & logistics	Factory, production, machinery, industry, logistics	Concrete topics on physical manufacturing and supply chains.
3	Digital transformation & e-commerce	E-commerce, digital, software, marketing, startup	Digitally focused themes like software development and online marketing.
4	Lifestyle, health, & culture	Health, lifestyle, culture, art, hobby	A smaller, more diverse cluster for health, arts, and other niche topics.

signals in video content are often distributed across multiple modalities, including spoken language, background audio, and visual cues. Exclusive reliance on textual metadata therefore risks overlooking salient information, particularly when titles are short, generic, or otherwise non-descriptive. To address this limitation, we envisage a multimodal extension in which SBERT-based textual embeddings are fused with visual representations (e.g., features extracted from key frames using CNNs or vision transformers) and audio-derived embeddings (e.g., from automatic speech recognition transcripts or prosodic features). Such a multimodal fusion layer would enable the system to disambiguate noisy or sparse textual descriptions and to capture thematic content that is conveyed primarily through imagery, speech, or other non-textual channels.

3) Clustering behavior and probabilistic topic modeling: in our research, k-means yields negative silhouette scores, while HDBSCAN assigns 100% of the samples to the noise class. Crucially, these outcomes are not merely artefacts of suboptimal parameter choices but stem from the intrinsic geometry of the SBERT embedding space for the StoryBox corpus. The 384-dimensional bilingual embeddings do not form compact, well-separated clusters; instead, they populate a continuous semantic manifold characterized by smooth transitions between overlapping themes such as education, entrepreneurship, industry, and technology. In such a setting, density-based methods like HDBSCAN tend to classify most points as noise, whereas centroid-based approaches like k-means impose artificially rigid partitions, naturally resulting in low cluster validity indices.

We therefore interpret these low scores not as a pure algorithmic failure but as reflective of the inherent thematic blurring and polysemy present in real-world multitopic video content. This behavior further underscores the need for more flexible topic modeling paradigms. We envision a promising direction for future work in incorporating probabilistic or soft clustering methods, such as BERTopic, hierarchical Dirichlet processes, and related Bayesian nonparametric models, and reporting side-by-side comparisons in terms of cluster consistency, stability, and downstream classification performance.

Future research should focus on multimodal integration to obtain a comprehensive perspective of video material. Visual characteristics, such as scene identification, object detection, and sentiment analysis, can decode contextual and emotional aspects included within frames. Audio analysis, including music genre classification, speech emotion recognition, and sound event detection, will uncover auditory nuances critical to interpretation. Furthermore, temporal analysis, through techniques such as shot boundary detection and narrative structure analysis, can reveal the narrative dynamics that shape the impact of a video. By combining these modalities, systems can achieve a richer and more comprehensive understanding of video content.

Architectural innovations are equally crucial to improve performance. Hierarchical classification, with taxonomies that span multiple levels of granularity, can allow for a more precise and flexible categorization of video content. Graph neural networks offer a promising approach by modeling relationships between videos as graph structures, capturing intricate dependencies across datasets.

Furthermore, continuous learning processes can ensure that models adapt to changing content trends without catastrophic forgetting, hence preserving relevance in dynamic situations. These developments will lead to more powerful and scalable video analysis systems.

To prove generality and practical applicability, extensive validation is necessary. Numerous datasets produced from multiple platforms, including YouTube, Vimeo, and TikTok, that reflect various types of content and user habits should be included in the evaluations. Addressing linguistic diversity and improving global accessibility can be achieved by implementing expanded language coverage, which includes more than 50 languages. Systems can maintain their success and relevance across multiple environments and time periods by demonstrating how content trends have evolved through longitudinal studies that examine topic progression over time.

Future studies could focus on creating unique theme clusters for emerging topics such as climate change, renewable energy, and the circular economy in video materials using models such as LDA and BERTopic. By creating comprehensive workflows for these processes using platforms such as KNIME [59], the security of sensitive data can be maintained throughout the entire process. Public debates, educational gaps, and this targeted analysis of innovative sustainability solutions are expected to provide valuable information for policy makers and the scientific community.

7. Conclusion

The research proposes a complete approach for automated bilingual (Turkish– English) theme classification of video content using multilingual transformer models, clustering algorithms, and LLM-based labeling. The framework integrates preprocessing, sentence-level embedding, zero-shot classification, clustering evaluation, and semantic consistency analysis into a coherent pipeline, supported by implementation details, hyperparameter settings, and practical insights for production deployment. Overall, the study demonstrates that bilingual processing can substantially improve thematic coherence over monolingual baselines for heterogeneous YouTube video collections, while simultaneously revealing the limitations of hard clustering in such fuzzy semantic spaces. Future work should therefore focus on multimodal integration (text, audio, and visual cues), probabilistic or soft topic models, and larger, genuinely multilingual datasets to enhance the robustness and adaptability of the proposed framework.

Acknowledgement

The authors thank the YouTube StoryBox team for providing access to their video collection for research purposes. Computational resources were provided by Google Colaboratory Research Credits Program. This study was produced within the scope of the project titled “Obtaining qualitative research themes and categories with Large Language Model on YouTube videos” (Project No: GAP-2024-593), supported by Bilecik Şeyh Edebali University Scientific Research Projects Unit.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/onder-ozturk/youtube-video-analyzer.

Author Contribution Statement

Hüseyin Parmaksız: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Önder Öztürk:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Osman Akarsu:** Conceptualization, Validation, Investigation, Resources, Writing – review & editing, Supervision, Project administration.

References

[1] Oh, E. S. (2022). The global digital content landscape. In A. Taubman (Ed.), *Trade in knowledge: Intellectual property, trade and development in a transformed global economy* (pp. 323–351). Cambridge University Press. https://doi.org/10.1017/9781108780919.013

[2] Nayak, A. A., & Dharmanna, L. (2020). A comprehensive survey on content analysis and its challenges. *Evolutionary Computing and Mobile Sustainable Networks*, *53*, 761–771. https://doi.org/10.1007/978-981-15-5258-8_70

[3] Singh, S., & Mahmood, A. (2021). The NLP cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access*, *9*, 68675–68702. https://doi.org/10.1109/ACCESS.2021.3077350

[4] Gera, A., Halfon, A., Shnarch, E., Perlitz, Y., Dor, L. E., & Slonim, N. (2022). Zero-shot text classification with self-training. In *Conference on Empirical Methods in Natural Language Processing*, 1107–1119. https://doi.org/10.18653/v1/2022.emnlp-main.73

[5] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2009). Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Transactions on Networking*, *17*(5), 1357–1370. https://doi.org/10.1109/TNET.2008.2011358

[6] Yousaf, K., & Nawaz, T. (2022). A deep learning-based approach for inappropriate content detection and classification of YouTube videos. *IEEE Access*, *10*, 16283–16298. https://doi.org/10.1109/ACCESS.2022.3147519

[7] Qiu, X., Shi, S., Li, B., Tan, X., Gao, Y., & Li, S. (2025). LLM-GAODE: Large-language-model augmented neural ordinary differential equation network for video nystagmography classification. *Knowledge-Based Systems*, 114050. https://doi.org/10.1016/j.knosys.2025.114050

[8] Mrinalini, K., Vijayalakshmi, P., & Nagarajan, T. (2022). SBSim: A Sentence-BERT similarity-based evaluation metric for Indian language neural machine translation systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1396–1406. https://doi.org/10.1109/TASLP.2022.3161160

[9] Alcoforado, A., Ferraz, T. P., Gerber, R., Bustos, E., Oliveira, A. S., Veloso, B. M., …, & Costa, A. H. R. (2022). ZeroBERTo: Leveraging zero-shot text classification by topic modeling. In *International Conference on Computational Processing of the Portuguese Language*, 125–136. https://doi.org/10.1007/978-3-030-98305-5_12

[10] Amjad, F., Khan, F., Tahir, S., Yaqoob, T., & Abbas, H. (2022). ENCVIDC: An innovative approach for encoded video content

classification. *Neural Computing and Applications*, *34*(21), 18685–18702. https://doi.org/10.1007/s00521-022-07480-2

[11] Teutsch, M., Sappa, A. D., & Hammoud, R. I. (2022). Detection, classification, and tracking. In M. Teutsch, A. D. Sappa, & R. I. Hammoud (Eds.), *Computer vision in the infrared spectrum* (pp. 35–58). Springer. https://doi.org/10.1007/978-3-031-01826-8_4

[12] Lindeberg, T. (2012). Scale invariant feature transform. *Scholarpedia*, *7*(5), 10491. http://dx.doi.org/10.4249/scholarpedia.10491

[13] Said, Y., Atri, M., & Tourki, R. (2011). Human detection based on integral histograms of oriented gradients and SVM. In *International Conference on Communications, Computing and Control Applications*, 1–5. https://doi.org/10.1109/CCCA.2011.6031422

[14] Liu, L., Pietikäinen, M., Qin, J., Ouyang, W., & Van Gool, L. (2020). Efficient visual recognition. *International Journal of Computer Vision*, *128*(8), 1997–2001. https://doi.org/10.1007/s11263-020-01351-w

[15] Xiao, L., Yan, Q., & Deng, S. (2017). Scene classification with improved AlexNet model. In *International Conference on Intelligent Systems and Knowledge Engineering*, 1–6. https://doi.org/10.1109/ISKE.2017.8258820

[16] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *IEEE International Conference on Computer Vision*, 4489–4497. https://doi.org/10.1109/ICCV.2015.510

[17] Kujani, T., & Kumar, V. D. (2023). Head movements for behavior recognition from real time video based on deep learning ConvNet transfer learning. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7047–7061. https://doi.org/10.1007/s12652-021-03558-2

[18] Huang, C., Fu, T., & Chen, H. (2010). Text-based video content classification for online video-sharing sites. *Journal of the American Society for Information Science and Technology*, *61*(5), 891–906. https://doi.org/10.1002/asi.21291

[19] Yadav, S. M., & Chaware, S. M. (2025). Detection and classification of objects in video content analysis using ensemble convolutional neural network model. *International Journal of Image and Graphics*, *25*(02), 2550006. https://doi.org/10.1142/S0219467825500068

[20] Duong, H. T., Le, V. T., & Hoang, V. T. (2023). Deep learning-based anomaly detection in video surveillance: A survey. *Sensors*, *23*(11), 5024. https://doi.org/10.3390/s23115024

[21] Bai, L., Lao, S., Jones, G. J., & Smeaton, A. F. (2007). Video semantic content analysis based on ontology. In *International Machine Vision and Image Processing Conference*, 117–124. https://doi.org/10.1109/IMVIP.2007.13

[22] Johnson, S. J., Murty, M. R., & Navakanth, I. (2024). A detailed review on word embedding techniques with emphasis on Word2Vec. *Multimedia Tools and Applications*, *83*(13), 37979–38007. https://doi.org/10.1007/s11042-023-17007-z

[23] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 1532–1543. https://doi.org/10.3115/V1/D14-1162

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., …, & Polosukhin, I. (2017). Attention is all you need. In *International Conference on Neural Information Processing Systems*, 6000–6010. https://doi.org/10.48550/arXiv.1706.03762

[25] Singhal, N., Yadav, A., Ankush, A., Singh, G., & Kumar, R. (2025). Leveraging XLM-RoBERTa with CNN and BiLSTM for Hinglish toxicity detection. *Journal of Communications Software and Systems*, *21*(4), 394–403. https://doi.org/10.24138/jcomss-2025-0133

[26] Ezugwu, A. E., Shukla, A. K., Agbaje, M. B., Oyelade, J., Alloghani, M., Gupta, A., & Gupta, D. (2021). Automatic clustering algorithms: A systematic review and bibliometric analysis of relevant literature. *Neural Computing and Applications*, *33*(12), 6247–6306. https://doi.org/10.1007/s00521-020-05395-4

[27] Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035. https://dl.acm.org/doi/10.5555/1283383.1283494

[28] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining*, 226–231. https://dl.acm.org/doi/10.5555/3001460.3001507

[29] Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

[30] Akarsu, O., & Parmaksız, H. (2025). Anatomy of digital leadership studies: An analysis with topic modeling approaches. *Business and Economics Research Journal*, *16*(2). https://doi.org/10.20409/berj.2025.463

[31] McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

[32] Zhou, Z. H. (2025). *Ensemble methods: Foundations and algorithms*. USA: CRC press.

[33] Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1), 1–39. https://doi.org/10.1007/S10462-009-9124-7

[34] Kadhim, R. I., Al-Mukhtar, F., Guron, A. T., Shwayaa, A. K., Al-Sharify, T. A., Al-Attar, B., …, & Hashim, W. A. (2024). Multimodal deep learning for video classification. In *International Symposium on Innovative Approaches in Smart Technologies*, 1–8. https://doi.org/10.1109/isas64331.2024.10845289

[35] Sharma, A. K., Chaurasia, S., & Srivastava, D. K. (2020). Sentimental short sentences classification by using CNN deep learning model with fine tuned Word2Vec. *Procedia Computer Science*, *167*, 1139–1147. https://doi.org/10.1016/j.procs.2020.03.416

[36] Gan, L., Teng, Z., Zhang, Y., Zhu, L., Wu, F., & Yang, Y. (2022). SemGloVe: Semantic co-occurrences for GloVe from BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2696–2704. https://doi.org/10.1109/TASLP.2022.3197316

[37] Ding, Y., Jia, M., Miao, Q., & Cao, Y. (2022). A novel time–frequency transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings. *Mechanical Systems and Signal Processing*, *168*, 108616. https://doi.org/10.1016/j.ymssp.2021.108616

[38] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of The Association for Computational Linguistics: Human Language Technologies*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[39] Sherif, A., & Sabty, C. (2024). Sentiment analysis for Egyptian Arabic–English code-switched data using traditional neural models and advanced language models. In *International Conference on Speech and Computer*, 54–69. https://doi.org/10.1007/978-3-031-78014-1_5

[40] Senthilselvi, A., Prawin, R. P., Harshit, V., Santhosh Kumar, R., & Senthil Pandi, S. (2024). Abstractive summarization of YouTube videos using lamini-flan-t5 llm. In *International Conference on Advances in Information Technology*, 1–5. https://doi.org/10.1109/ICAIT61638.2024.10690747

[41] Hu, C., Sun, X., Dai, H., Zhang, H., & Liu, H. (2023). Research on log anomaly detection based on Sentence-BERT. *Electronics*, *12*(17), 3580. https://doi.org/10.3390/electronics12173580

[42] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., …, & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of The Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[43] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

[44] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[45] Nanga, S., Bawah, A. T., Acquaye, B. A., Billa, M.-I., Baeta, F. D., Odai, N. A., …, & Nsiah, A. D. (2021). Review of dimension reduction methods. *Journal of Data Analysis and Information Processing*, *9*(3), 189–231. https://doi.org/10.4236/jdaip.2021.93013

[46] Ruffino, S., Karunaratne, G., Hersche, M., Benini, L., Sebastian, A., & Rahimi, A. (2024). Zero-shot classification using hyperdimensional computing. In *Design, Automation & Test in Europe Conference & Exhibition*, 1–2. https://doi.org/10.23919/DATE58400.2024.10546605

[47] Bao, C. (2021). K-means clustering algorithm: A brief review. *Academic Journal of Computing & Information Science*, *4*(5), 37–40. https://doi.org/10.25236/AJCIS.2021.040506

[48] Kumar, P., & Agrawal, R. (2024). Performance benchmarking of clustering methods using MNIST data. In *International Conference on IoT Based Control Networks and Intelligent Systems*, 39–44. https://ieeexplore.ieee.org/abstract/document/10823372

[49] Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

[50] Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies–Bouldin index for clustering evaluation. *Neurocomputing*, *528*, 178–199. https://doi.org/10.1016/j.neucom.2023.01.043

[51] Amin, M. F. I., Watanobe, Y., Rahman, M. M., & Shirafuji, A. (2025). Source code error understanding using BERT for multi-label classification. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3525061

[52] Chadha, A., & Kaushik, B. (2022). A hybrid deep learning model using grid search and cross-validation for effective classification and prediction of suicidal ideation from social network data. *New Generation Computing*, *40*, 889–914. https://doi.org/10.1007/s00354-022-00191-1

[53] Liu, W., Fu, X., & Strube, M. (2023). Modeling structural similarities between documents for coherence assessment with graph convolutional networks. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2306.06472

[54] Zhang, Q., Yang, Z., Huang, Y., Chen, Z., Cai, Z., Wang, K., …, & Gao, J. (2023). Enhancing model performance in multilingual information retrieval with comprehensive data engineering techniques. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2302.07010

[55] Horbach, A., Pehlke, J., Laarmann-Quante, R., & Ding, Y. (2023). Crosslingual content scoring in five languages using machine-translation and multilingual transformer models. *International Journal of Artificial Intelligence in Education*, 1–27. https://doi.org/10.1007/s40593-023-00370-1

[56] Akarsu, O. (2024). Girişimcilerin nedenini ve nasılını snlamak: "StoryBox" üzerinden bir çözümleme [Why and how of entrepreneurs: An analysis via storybox]. *Sosyal Mucit Academic Review*, *5*(1), 50–93. https://doi.org/10.54733/smar.1404649

[57] Ormiston, J., & Thompson, N. A. (2021). Viewing entrepreneurship "in motion": Exploring current uses and future possibilities of video-based entrepreneurship research. *Journal of Small Business Management*, *59*(5), 976–1011. https://doi.org/10.1080/00472778.2020.1866184

[58] Çoklar, A. N., & Cihangir, H. H. (2021). Using YouTube as an education environment: Examining follower views. *International Technology and Education Journal*, *5*(1), 50–60. https://dergipark.org.tr/en/pub/itej/article/1233250

[59] Shen, J. C., Su, N. J., & Lin, Y. B. (2025). Effective multi-class sentiment analysis using fine-tuned large language model with KNIME analytics platform. *Systems*, *13*(7), 523. https://doi.org/10.3390/systems13070523

How to Cite: Parmaksız, H., Öztürk, Ö., & Akarsu, O. (2026). Machine Learning-Based Theme Classification for Video Content Analysis: A Bilingual Approach on the StoryBox. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62026942