



Decoding content taxonomy in video for industry-specific contextual advertising

Waruna De Silva¹ · Anil Fernando¹

Received: 29 May 2024 / Revised: 30 September 2025 / Accepted: 28 October 2025
© The Author(s) 2026

Abstract

The exponential growth of free video platforms is redefining content consumption and its advertising dynamics. Beyond this, the progressive irrelevance of cookies really underlines the necessity for finding a genuinely new and proper way of recording and analyzing user behavior, with an implication for putting user privacy first by design, compliant with data protection and regulatory guidelines. As a result, contextual advertising has become a suitable strategy for the delivery of relevant and personalized advertisements to users. In light of this evolving process, this article proposes a novel framework to facilitate the discovery of contextual information on video content within the industrial context. The proposed framework seamlessly integrates the visual and audio features that are extracted from video content to obtain a comprehensive comprehension of videos. This method optimizes the presentation of ads within a context using a combination of multimodal analysis, industry taxonomy, and contextual advertising. The effectiveness of the framework is validated through experimental results using the YouTube-8M data set, demonstrating its potential to revolutionize contextual advertising by capturing the essence of video content and aligning it with the industry content taxonomy.

Keywords Natural language processing · Video contextual advertisements · Multimodal · Topic modelling · Content taxonomy standards · BERTopic · BERT

1 Introduction

The advertising digital ecosystem has experienced tremendous changes with the rise of free video streaming platforms that deliver hundreds of programs each day [1]. In addition, the transition to a cookie-free world presents a distinctive and urgent challenge. Especially so

✉ Waruna De Silva
orthalange.de-silva@strath.ac.uk

Anil Fernando
anil.fernando@strath.ac.uk

¹ Department of Computer and Information Sciences, University of Strathclyde, 16 Richmond Street, Glasgow G1 1XQ, UK

with respect to the access of behavioral data [2]. With cookies becoming obsolete, the ability to capture and analyze user behavior is one of the major concerns for business players. These bring into focus the need for integrating data analysis requirements with user privacy needs and evolving data protection regulation needs. While much is laid on the standards required to make effective ad requests across varied systems, it is seen that the requirement of relevant and personalized ad targeting practices is more needed than ever in this environment [3]. In addition, there are also the brand safety and user experience challenges mentioned earlier. Such challenges have highlighted the importance of context analysis in identifying suitable advertisements [4–6]. Therefore, exploring contextual advertising is a proactive point of view to overcome these issues and also a technique to deliver successful targeted advertising of videos in the industrial setting [3].

This paper introduces a contextual advertising framework for video platforms and is especially valuable in cold-start situations where the ad-delivery operation begins with new users or content without prior data. Traditional video advertising platforms often rely solely on video metadata [7–9]. However, the importance of assembling a relevant advertising selection using this method can result in suboptimal results. This is because the metadata are not always complete or comprehensive. The demand to implement contextual video advertising arises from the complexity of videos and TV categories, where understanding the broader picture goes beyond just metadata. Unlike other media, such as audio or text, video content benefits from the inclusion of visual elements. These visual aspects of videos are often underused. And this represents a critical oversight in contextual advertising platforms, as alignment with video content is critical.

As we explore the contents of this paper, our primary focus is towards the collection of diverse audio and visual multimodal data within video content. And then effectively fusion [10] them to reflect a thorough contextual sense of the video. This method enables a more complete understanding of contextual information. It is important to note that there has also been tremendous advancement in different research domains regarding visual and auditory modalities in videos. In the visual domain, enormous gains are there in some prime areas of research. This is particularly noted in the enhancement of contextual awareness in computer visions [11–14]. At the same time, in the field of audio, continuous stress is placed on the task of Automated Audio Captioning (AAC), which involves the creation of coherent sentences in natural language [15, 16]. It is also remarked that audio signals and their constituents, basically speech and ambient noise, are very important in the course of analysis. This is especially noted in the context of robust Automatic Speech Recognition (ASR) [17] and transcript identification.

It is important to consider in the multimodal analysis of video media that the capacity for information and the level of noise can be different from one modality of data to the other, as illustrated in Fig. 1. Each of the modalities gives unique and complementary insight into the underlying content. As such, with a realisation of the variation videos bring in their content, both auditive and visual, our method will try to go deeper into the way these sources rich in information can be put to full use in overcoming the constraints that are set by the lack of some other features. Having a depth of information capacity helps to correctly interpret the message of a video, and this is important to achieve correct targeting of the advertisements.

Our work addresses the problem of managing diverse multimodal data and the task of handling varying information and noise. We aim to use video topic recovery in this model with the intent of understanding the underlying themes of the video and going beyond video

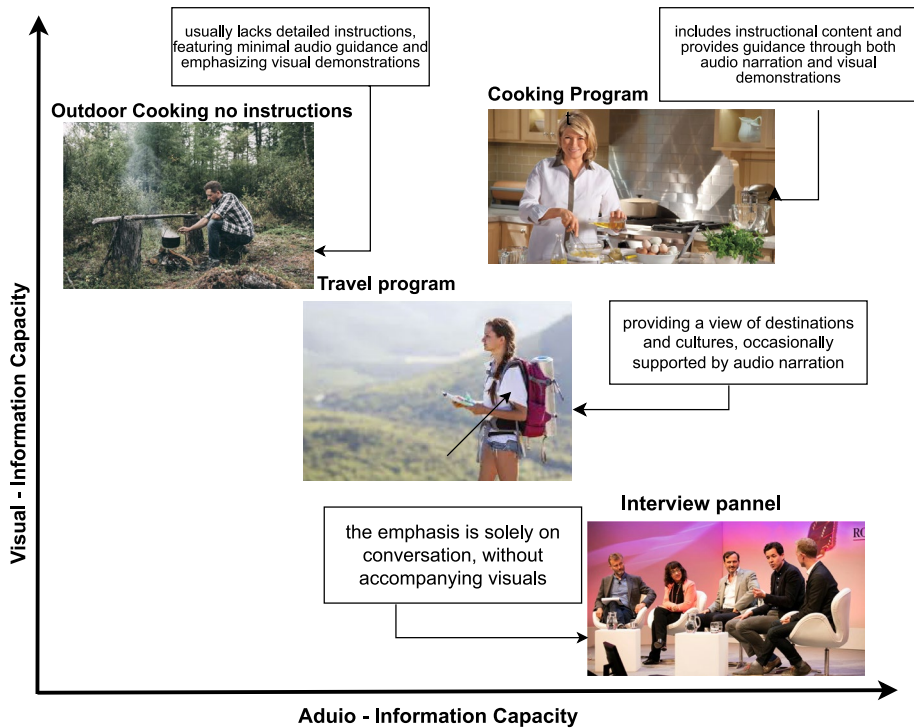


Fig. 1 Understanding Information Capacity in Diverse Data Modalities

classification. Our approach uses topic modelling [18] for the extraction of latent semantic units and themes in a video. Hence, it classifies its meaning by the extraction of subtopics, patterns, and the relationship between the information, which may not be immediately recognised. Topic recovery in one modality is limited to the naturally occurring constraints in that single modality. However, multimodal fusion [10] gives access to a lot more diverse information and hence implies an overall more accurate and robustness of the topic recovery process. Moreover, it should also be noted that the use of audiovisual data for this purpose adds to the richness of the domain-specific features of these topics.

Furthermore, the study is in the context of recovering topic labelling using an industry-specific content taxonomy [19]. Since unsupervised topic modelling does not provide external labels, further refinement of the labelling process was necessary [20]. Although traditional methods such as pointwise mutual information [21] have been effectively used for cluster labelling, our approach maps the identified video topics to the content taxonomy [19].

To validate the effectiveness of our framework, we present experimental results based on a YouTube-8M [22] sample data set. These evaluations analyse each component of the framework, its overall performance, and exhibit its potential to revolutionise the fields of contextual advertising.

1.1 Main contributions

We consolidate our contributions into three points:

1. **Unified multimodal framework.** We present a practical pipeline that fuses audio (ASR transcripts) and visual (Bag-of-Visual-Words) signals and adapts BERTopic for video, recovering semantically rich themes beyond metadata-only approaches.
2. **Standards-aligned taxonomy mapping.** We introduce a similarity-based mapping from discovered topics to industry taxonomies (e.g., IAB) using sentence embeddings with topic/term intensity weighting, yielding interpretable labels with confidence scores for ad targeting.
3. **Empirical validation at scale.** Using a YouTube-8M subset, we show high Top-5 success and that late fusion outperforms early fusion, evidencing scalability and deployment readiness.

The rest of the paper is structured as follows. In Section 2 we summarize related work on video understanding in the context of contextual advertising. In Section 3 we provide the details of our proposed methodology with the multimodal feature extraction and topic modeling framework. In Section 4 we report our experimental design and results, followed by a discussion in Section 5. Lastly, Section 6 concludes the paper and suggests areas for possible future work.

2 Related work

2.1 Metadata-based approaches for video advertising

Earlier work on contextual video advertising has proposed a system named VideoSense [7]. The recommendation system of VideoSense comprised title, tags, and queries and local visual-aural features such as color, movement, and audio. In yet another work [9], it was proposed to employ a process that applies five classes of video metadata to extract informative advertisements with a principal objective of reducing resource-hungry imagery and video processing operations. The work suggests a system in which the attributes of the metadata are weighted based on relevance and decisions taken based on a model of importances of blocks. In the same category of work, the work of SalAd [8] suggests a new advertising paradigm using textual metadata to assemble a candidate set of informative advertisements a prevalent paradigm in earlier work. SalAd further optimizes the process using webpage saliency to decrease deliberate ignoring of advertisements. A key drawback of existing studies lies in the fact that they do not conduct extensive video content analysis to extract the required contextual subtleties. The majority of existing methodologies critically depend on the source video's metadata or particular visual/audio features, usually unable to extract the complete semantic richness of video content. This drawback manifests critically in the situations during which the metadata proves incomplete, inconsistent, or inaccurately representative of the content.

2.2 Deep learning models for video content analysis

Building on earlier metadata-driven solutions, researchers later investigated deep learning methods for the more complex analysis of video. Some works looked at the specific detail of inserting advertisements into videos associated with objects. For instance, a deep CNN

structure [23] developed a system to detect objects in videos in order to make video commercials less intrusive. This work has limited generality for the insertion of commercial advertisements that are not directly associated with tangible objects. Another work [24] employed deep neural networks to detect the human features of bodies, poses, faces, and garments. Although useful for the advertising of clothing, this work has been admitted to be limited in its generality of applicability to other types of unconstrained video like movies or sport videos. More progress on deep learning for video advertising was introduced in [25], in which a unique CNN model proposed an efficient solution for content-targeted online video advertising. This solution is based on matching videos with advertisements using the former's corresponding semantic descriptions. Even though deep CNNs have achieved good results in numerous applications, application on video analysis is minimal considering the linear progress inherent in videos. In order to extract longer-range temporal patterns, the use of Recurrent Neural Networks (RNNs) has been considered [26]. Current techniques remain limited to small-scale datasets only. Combining the CNN and RNN models for end-to-end training on large-scale datasets remains difficult, especially on datasets like YouTube-8M.

2.3 Multimodal fusion techniques

As a solution to the deficiencies of the unimodal techniques, the multimodal fusion techniques for video commercials have been explored. In [27], the framework was introduced to extract descriptive comprehensive features from various modalities. The approach considered the features such as location, emotion, objects, audio, and topics applied to videos and commercials. These various representations are blended together to create a single coherent representation, simplifying the process of scene matching. However, this approach lacks a thorough examination of the contribution of each modality to the most resilient semantic modality for contextual understanding. Despite these, most research works focus on contextual understanding of local scenes in videos to derive ad-insertion points within specific contexts [28, 29]. While relevant in some domain or niche area specifications, these approaches become difficult to scale to a wider spectrum of video content. This limitation highlights the need for careful analysis of hidden themes and overall themes of videos, which would enable advertisers and downstream applications to make informed decisions and deliver the right ads with the right content.

2.4 Topic modeling in video context

Video thematic content identification has become a key element for video contextual advertising [30]. Classical topic modeling methods have been used for identifying hidden themes from video content, but the methods suffer from several critical issues in handling multimodal content [31]. While topic modeling of text is long familiar, uses of these techniques with other modalities such as visuals and audio require custom adaptations [32]. Merging topic modeling with industry taxonomies is a particular underexplored area of video ads research. The discovered topics need to be framed such that they can be interpreted by downstream applications and advertisers and can coexist with popular industry taxonomies [19]. This aspect, as far as our understanding is concerned with video advertising, has not been fully explored in previous research relating to video advertising. The lack of robust

methods for mapping video content to standardized taxonomies represents a significant gap in the literature.

2.5 Recent advances: video understanding

Recent advances in self-supervised learning and Transformer architectures have raised the prospects of video understanding to interesting possibilities [33]. Self-supervised learning techniques have reached all-time record-breaking performance in learning rich video representations with minimal or without labelled data. Contrastive learning methods such as MoCo [34], SimCLR [35], and CLIP [36] originally developed for image-text alignment but influential in shaping multimodal contrastive frameworks—have been adapted to video information and enabled the learning of temporal dynamics and semantic relationships that in some scenarios surpass supervised methods. Masked autoencoding approaches (e.g., MotionMAE [37], MGMAE [38], ViC-MAE [39]) further enhance this by modeling both spatial appearance and temporal motion patterns in a self-supervised fashion. Transformer architectures, in an analogous manner, have reformed the horizon of video understanding by allowing the modelling of long-range dependencies of spatio-temporal information. Models such as VideoBERT [40] have evidenced the superiority of joint video-text representation learning, and models such as TimeSformer [41] and ViViT [42] have reached dazzling performance in spatio-temporally efficient modelling. Such developments hold immense promise in aiding the improvement of contextual systems of advertising through more complete understanding of video information. Beyond these trends, transformer-based salient object detection (SOD) explicitly models global–local relations with compensative fusion; the Collaborative Compensative Transformer Network (CCTNet) alternates collaborative relation modelling with compensative fusion to improve saliency prediction [43]. Complementary to our privacy-by-design motivation, FedMDD calibrates global decision boundaries for federated long-tailed settings via multi-deliberation post-hoc calibration and local–global feature-contrast constraints; although orthogonal to our modelling pipeline, these ideas are pertinent for future privacy-preserving, cross-partner deployments and for handling skewed taxonomy-label frequencies in decentralised data [44].

Despite all these advancements, hardly any have been used in contextual advertising due to the fact they are not specifically optimized for long-range video understanding. The integration of transformer models and self-supervised models with the advertising industry presents a major research avenue that has been hardly touched, in part due to higher compute costs and scalability [33]. These methods have the ability to remedy a vast number of inadequacies in the existing techniques, most particularly the identification of complex thematic content and its conversion into common advertising categories.

2.6 Summary and research gaps

Earlier studies on contextual video ads revealed some limitations. Metadata-based approaches often cannot cover all the semantic richness of video content, with incomplete or imprecise context. The deep learning approach holds potential but is limited by scaling and data set deficiencies. Different multimodal fusion techniques are proposed but lack robust tools for estimating the contribution of disparate modalities and thus exhibit brittleness for problematic situations. Topic modeling has been used throughout the literature for

the vast majority of text-based domains but not for multimodal video data. Finally, while great strides have been taken in self-supervised learning and Transformers, these methods have not widely been applied to ad applications due to prohibitive computational costs and scarcity of optimization for long-distance video understanding.

These gaps suggest the need for a paradigm which goes beyond unimodal or metadata-based techniques, effectively utilizes multimodal signals, and harmonizes extracted knowledge with the practices of the industrial communities.

2.7 Technical challenges

Translating the research gaps described above into practice holds several technical challenges for which our framework proposes solutions:

1. **Incomplete and unreliable metadata.** Existing systems rely heavily on titles, tags, and descriptions, which are often sparse, inconsistent, or misleading.
2. **Multimodal complexity.** Video combines a visual and an audio stream, each carrying an amount of semantic information and noise. Extracting, aligning, and synthesizing these multimodal sources into a single representation does not come easily.
3. **Adapting topic modeling to video.** Classic topic models are text-based. Their extension to video involves projecting visual features into textual-like features and integrating them with audio transcripts while retaining semantic abundance.
4. **Mapping to standardized taxonomies.** Latent topics from unsupervised models cannot be directly interpreted for ad apps. Robust mapping onto traditional taxonomies (e.g., IAB) is necessary for deployment but is technically difficult.
5. **Scalability and domain adaptation.** Efficient systems must process large data sets (e.g., the YouTube-8M) without forgetting detailed, domain-rich topics. Striking a balance between computational resources and contextual correctness is a large problem.

These challenges form the foundation for the methodology detailed during Section 3, which provides our multimodal topic modeling framework and approach for mapping the taxonomy.

3 Proposed methodology

To address the technical challenges outlined in Section 2.7, we formalize the problem and present a multimodal topic modeling and taxonomy-mapping framework.

3.1 Problem definition

We consider a video $V = \{F, A\}$ with visual frames F and an accompanying audio track A . The objective is to infer a set of industry-standard content taxonomy labels \mathcal{Y} (e.g., IAB) and associated confidence scores $s \in [0, 1]^{|\mathcal{Y}|}$ that reflect the primary and secondary themes present in V and are suitable for downstream ad targeting. Concretely, we extract visual features and transform them into a Bag-of-Visual-Words representation x_v , and we transcribe the audio into text x_a via ASR. A multimodal encoder produces topic representations T with

weights w from (x_v, x_a) ; a taxonomy-mapping function then aligns T to \mathcal{Y} using embedding-based similarity with topic/term intensity weighting, yielding ranked labels and scores:

$$(x_v, x_a) \xrightarrow{\text{encoder}} T, w \quad \text{and} \quad (T, w) \xrightarrow{\text{taxonomy map}} \{(y, s_y)\}_{y \in \mathcal{Y}}.$$

Our formulation explicitly acknowledges the multimodal and temporal nature of video (heterogeneous information capacity and noise across modalities; variable quality in resolution, lighting, and acoustics) and the requirement that outputs be interpretable in standardized taxonomies. Evaluation focuses on accuracy at top- k , coverage of relevant labels, and ranking quality, with robustness to modality imbalance and scalability to large corpora detailed in Section 3.2.

3.2 Solution overview

Figure 2 provides the flowchart of the proposed method, showing the end-to-end pipeline from multimodal feature extraction to taxonomy-aligned outputs. We start with the extraction of video and audio features, respectively, from Amazon Rekognition [45] and OpenAI/Whisper[46]. Recently, late- and early-fusion methods have been able to concatenate different modalities of audio and video [47]. In late fusion adaptation, we individually apply topic modelling to these modalities to get contextual representation. In early fusion adaptation, the original features from audio and video are early fused for direct topic modelling, and the representations are synchronised with industry taxonomies.

Our underlying assumption is that by recommending prominent topic terms from modelled topics that are in close alignment with industry context taxonomy standards, we can establish a standardized method for determining advertising categories. Meaning that after deriving the topics from topic modelling and modality fusion, we add those topics in a vector space along with industry context taxonomy standards as vector embeddings. Then we establish advertising categories by considering the topic terms that are close to each taxonomy standard. This, established on the basis of standards, facilitates effective communication with integrated systems within the digital advertising ecosystem. It also allows various components to interact seamlessly during the exchange of ad requests. Fundamentally, this process is supported by two core components that are inherent to our framework, video contextual representation, and its integration with industry-specific context taxonomies.

In our methodology, we understand that topic modelling methods help in understanding themes in data and are designed mainly for textual datasets [48]. These designs prioritise

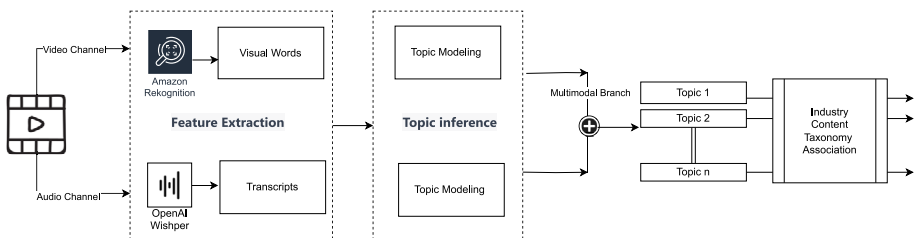


Fig. 2 Flowchart of the proposed method (system overview). Visual and audio features are extracted, topics are inferred, modalities are fused, and topics are mapped to standardized taxonomies to produce ranked labels

the meaning and context of words. Although audio transcripts prove effective for audio information, challenges have arisen when applying these models to conventional features in computer vision. Notably, a key objective in structuring this information is to effectively translate visual content into precise and concise textual descriptions. Therefore, our methodology embraces a concept similar to text mining, specifically adapting the Bag-of-Words(BoW) model into a Bag-of-Features(BoF) framework [49] tailored for visual information analysis. The BoF technique has been widely used in numerous computer vision and machine learning competitions [50, 51].

One of the key characteristics of the framework is the conversion of visual features into a Bag of Visual Words (BoVW) format [52]. This transformation is an important procedural step in which image features are treated as words. This in turn allows for a Bag-of-Words approach in computer vision applications and also serves to augment data interoperability. This ensures a seamless integration of findings derived from both audio and visual data sources.

We have used Amazon Rekognition [45] to pull the BoF from the video and then convert them into a BoVW. This collaboration with external services is our way of combining advanced feature extraction methods with established techniques for effective video analysis. Representing these extracted features in a standardised format, we formulate a powerful representation of the BoVW, encapsulating the rich visual information.

To identify essential audio features, we rely on Whisper [46], an advanced OpenAI ASR system. Whisper's sophisticated ASR capabilities enable us to transform spoken language into accurate written text with extraordinary efficiency.

3.3 Topic modelling

A topic model is used to capture the essence of video content. This approach is significant in that it requires a thorough understanding of all contextual facets within the video. It also facilitates the identification of numerous advertising opportunities embedded in the content. Topic models [53, 54], are effective in automatically identifying topics from features. It commonly uses a BoW technique that quantifies the presence of words by studying occurrences and correlations [55, 56].

Building on the foundation of our topic modelling approach, we use BERTopic [57] for our model because it increasingly incorporates neural components [58]. This choice allows us to utilise pre-trained models on a large corpus, such as BERT iterations [59]. Our choice was guided by the distinct advantages that BERTopic offers over traditional topic modelling techniques such as LDA [60]. BERTopic provides a semantic representation for each term and effectively addresses the problem of mismatching the vocabulary [57]. Furthermore, BERTopic differentiates itself with its ability to generate more easily interpretable topics, support multilingual analysis, and autonomously determine the optimal number of topics [61]. Compared to dynamic LDA [60], BERTopic requires considerably less fine-tuning of the hyperparameters, thus simplifying the manual tasks involved in the process [61]. BERTopic also introduces the concept of noise topic, effectively preventing unrelated documents from being mistakenly assigned to other topics. This enhancement greatly improves the quality of topic representations. The technical capability of BERTopic covers the identification of dynamic topic modelling with reduced subjectivity. This was made evident by previous research [62].

BERTopic’s modular pipeline that aims to make document topics interpretable. First, the documents—in our case, audio transcripts and visual features—are passed through an encoder that generates document embeddings using powerful pre-trained transformer-based language models. Next, it performs dimensionality reduction techniques using Uniform Manifold Approximation and Projection [63] and clustering through Hierarchical Density-Based Spatial Clustering of Applications with Noise [63] to generate clusters, which contain semantically similar documents. Finally, it constructs the topic representations using a class-based term frequency-inverse document frequency term. In practice, this usually involves finding the *class-based Term Frequency–Inverse Document Frequency (c-TF-IDF)* [57] of a word, which is the frequency of the word in the topic, divided by the total number of words.

In addition, it divides the average number of words in each topic by the total frequency of that word across all topics. This holistic method ensures the generation of coherent and relevant topics. Meanwhile, it also takes into account considerations related to word frequency and topic representation extension.

$$W_{t,c} = \text{tf}_{t,c} \cdot \log \left(1 + \frac{A}{\text{tf}_t} \right) \quad (1)$$

In BERTopic, after documents are embedded and clustered, c-TF-IDF (1) is used to create a class-based representation of topics by transforming the textual content of each topic cluster into a meaningful topic vector.

- $\text{tf}_{t,c}$: Term frequency of term t in class/topic c (i.e., how often word t appears in documents belonging to topic c).
- tf_t : Total frequency of term t across all classes.
- A : Total number of all words in all classes (i.e., the total length of the corpus).
- $W_{t,c}$: Importance (or weight) of term t in class/topic c .

3.4 Topic model training

We employ a non-traditional approach to address the difficulties caused by constrained vocabulary and context in BoVW representations [64]. Our approach uses a large textual corpus in conjunction with pre-trained BERT-based models, instead of training directly with video object labels.

This corpus, rich in descriptive content, forms the foundation of our methodology. By fine-tuning a pretrained BERTopic model on this textual corpus, we enhance its grasp of language semantics, context, and nuances. This refined model is then applied to object labels extracted from video frames, employing semantic similarity to cluster labels into topics. Our alternative method shows to be more contextually aware and linguistically adept compared to traditional approaches [65].

Our approach copes successfully with the limitations of direct training in object labels, and thus demonstrates feasibility. Using a pre-trained BERTopic model for video topic analysis brings some advantages. Our novel approach not only overcomes these issues, but it shows the way forward in directions that can enable more nuanced and context-rich video topic analysis.

3.5 Multimodal topic representation

In the final stage, we integrate multimodal fusion techniques with proper consideration of both topic probability (P_{tw}) and topic term probability (P_{ttw}) during the process of mapping taxonomies. The fusion and mapping process has been detailed through Fig. 3, materializing the Adjusted Similarity Score (ASS) of (4) under both early- and late-fusion frameworks. Success score calculation favors the topic with the highest probability and assigns prominence to topic terms under highlighted topics during the mapping process of taxonomies. The notion of ‘topic intensity’, as introduced by [66], aims at a similar objective as ours of quantifying the relevance and significance of topics.

Furthermore, the integration of advanced vector space embeddings into the process of mapping similarity between topic terms and taxonomy terms is based on the theoretical framework of semantic representation. Vector space embeddings convert words into contin-

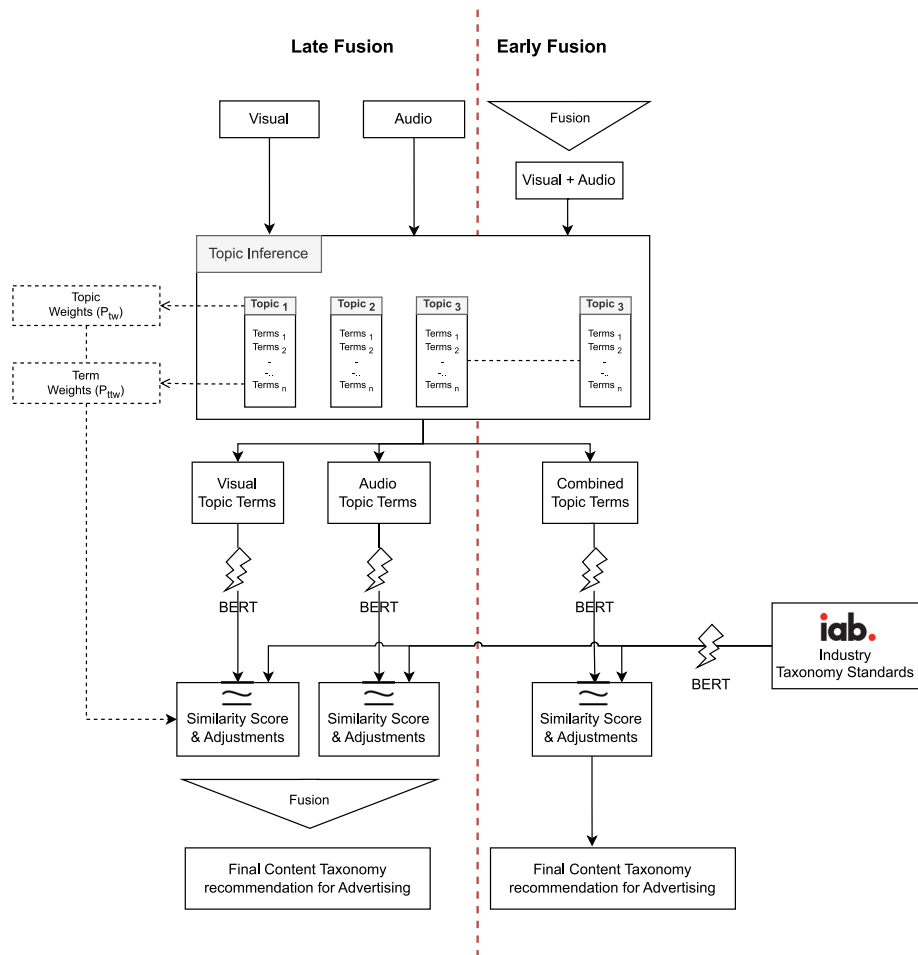


Fig. 3 Detail of the multimodal fusion and intensity-weighted taxonomy mapping (corresponding to the central block in Fig. 2). It illustrates early/late fusion and the Adjusted Similarity Score (ASS) used for taxonomy mapping

uous vector representations, where the spatial proximity represents semantic associations. The importance of utilising advanced vector space embeddings enables a detailed and contextually comprehensive comprehension of the semantic similarity between topic terms and taxonomy terms.

We use pre-trained neural embeddings generated using the BERT [59] sentence transformer. Specifically, we create neural embeddings for Contextual Taxonomy Terms (EMB_C), visual topic terms (EMB_V), and Audio Topic Terms (EMB_A). These embeddings enable us to quantify the semantic similarity between content taxonomy keywords and topic keywords, and thus mapping them to a 384-dimensional dense vector space. Figure 4 zooms into the embedding-based similarity sub-module.

Our primary objective is to quantify similarity scores ($SS_{C,V}$ and $SS_{C,A}$) that indicate the level of correspondence between these categories. Calculating $SS_{C,V}$ involves evaluating the cosine similarity between embeddings for Contextual Taxonomy terms (EMB_C) and Video Topic Terms (EMB_V), while $SS_{C,A}$ involves evaluating the cosine similarity between (EMB_C) and Audio Topic Terms (EMB_A). We employ cosine similarity as a scoring method to identify embedded visual/audio topic terms that have semantic similarity with taxonomy context terms.

The cosine similarity formulas are given by:(2),(3)

$$SS_{C,V} = \frac{EMB_C \cdot EMB_V}{\|EMB_C\| \cdot \|EMB_V\|} \tag{2}$$

$$SS_{C,A} = \frac{EMB_C \cdot EMB_A}{\|EMB_C\| \cdot \|EMB_A\|} \tag{3}$$

Subsequently, we refine the similarity score (SS) by multiplying it by the probability of the relevant topic probability (P_{tw}) associated with the specific topic and the respective term probability (P_{ttw}). This adjustment ensures that the SS reflects the topic-specific significance, enabling a more context-aware evaluation. We refer to these attentions because

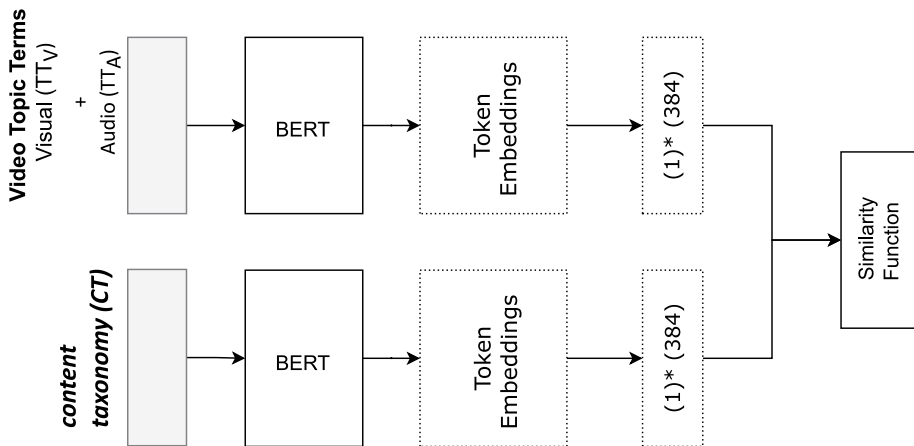


Fig. 4 Zoom-in on the embedding-based similarity computation between topic terms and taxonomy terms (the scoring sub-module inside Fig. 3); it implements the cosine similarities in Eqs. (2)–(3)

the topic and topic term intensity aim to identify the most relevant SS for associating the taxonomy with the topic.

The formulas for the adjusted similarity score (ASS) are given by (4)

$$ASS = SS \times P_{tw} \times P_{ttw} \quad (4)$$

The final result of this process is the adjusted similarity score (ASS), which contains the semantic congruence between the content taxonomy keywords and the topic keywords while accounting for the topic-specific relevance.

Statistical significance of multimodal fusion results was validated using paired Wilcoxon signed-rank tests ($\alpha = 0.05$), addressing potential dataset bias concerns through rigorous hypothesis testing.

3.6 Algorithm implementation

To realize the above explained multimodal fusion and taxonomy mapping framework, we provide here an Algorithm 1. The pseudocode encapsulates the end-to-end process from taxonomy ranking at the feature level down to the final taxonomy ranking, applying both early and late fusion schemes with modified similarity scoring. The above algorithm acts as an implementation roadmap of our experimental verification presented in Section 4.

Algorithm 1 Contextual Video Taxonomy Mapping via Multimodal Topic Modeling.

Require: Video $V = \{F, A\}$ ▷ F : Frames, A : Audio
Require: Industry Taxonomy Terms T_{taxonomy}
Require: FusionType $\in \{\text{'Late'}$, 'Early' $\}$
Ensure: Ranked Taxonomy Labels with Confidence Scores
1: $BoVW \leftarrow \text{ExtractVisualFeatures}(F)$ ▷ e.g., Amazon Rekognition
2: $ASR_Text \leftarrow \text{TranscribeAudio}(A)$ ▷ e.g., OpenAI Whisper
3: **if** $FusionType == \text{'Early'}$ **then**
4: $Fused_Input \leftarrow \text{Concatenate}(BoVW, ASR_Text)$
5: $Topic_List \leftarrow \text{BERTopic}(Fused_Input)$
6: **else**
7: $Visual_Topics \leftarrow \text{BERTopic}(BoVW)$
8: $Audio_Topics \leftarrow \text{BERTopic}(ASR_Text)$
9: $Topic_List \leftarrow Visual_Topics \cup Audio_Topics$
10: **end if**
11: $T_terms \leftarrow \text{ExtractTopTerms}(Topic_List)$
12: $EMB_Topics \leftarrow \text{BERT_Embed}(T_terms)$
13: $EMB_Taxonomy \leftarrow \text{BERT_Embed}(T_{\text{taxonomy}})$
14: **for** each taxonomy term t in T_{taxonomy} **do**
15: **for** each topic term w in T_terms **do**
16: $SSC[t][w] \leftarrow \text{CosineSimilarity}(EMB_Taxonomy[t], EMB_Topics[w])$
17: $ASS[t][w] \leftarrow SSC[t][w] \times P_{tw} \times P_{ttw}$
18: **end for**
19: $ASS_Total[t] \leftarrow \text{Aggregate}(ASS[t][:])$ ▷ e.g., weighted sum
20: **end for**
21: $Ranked_Labels \leftarrow \text{SortDescending}(ASS_Total)$
22: **return** Top-N $Ranked_Labels$ with Scores

3.7 Implementation details

To facilitate reproducibility and clarify engineering choices, we summarise the end-to-end setup in Table 1. The table is organised by pipeline stage—environment, preprocessing, feature extraction for the visual/audio streams, BERTopic configuration, fusion procedure, taxonomy mapping, training data, and evaluation protocol—so readers can see default settings at a glance. Unless otherwise stated in the ablations, all experiments use these defaults. We use a shared sentence-transformer encoder (384-d) for both topic modelling and taxonomy

Table 1 Implementation summary of the proposed framework

Component	Details
Environment	Amazon SageMaker <code>ml.t3.xlarge</code> (CPU-only; 4 vCPU, 16 GiB RAM), Amazon Linux 2; Python 3.10. Key libs: <code>sentence-transformers</code> , <code>umap-learn</code> , <code>hdbscan</code> , <code>bertopic</code> , <code>numpy/scipy</code> , <code>whisper</code> , <code>boto3</code> .
Data preprocessing	<ul style="list-style-type: none"> •Video: decode and frame export via <code>ffmpeg</code>; scene segmentation with <code>PySceneDetect</code> (<i>content</i> detector); extract <i>one keyframe per detected scene</i>; resize keyframes to shorter side 256 px (preserve aspect ratio). •Audio: resample to 16 kHz mono for ASR; enable auto language detection when applicable.
Visual features (BoVW)	<ul style="list-style-type: none"> • Run AWS Rekognition <i>only on keyframes</i>; retain top-$K=10$ labels with confidence ≥ 0.55. • Normalise labels (lowercase, lemmatise, synonym map) and aggregate per-video to form a Bag-of-Visual-Words.
Audio features (ASR)	Whisper transcripts; normalise text (lowercase, remove non-linguistic tokens; light punctuation/number filtering).
Topic modelling (BERTopic)	Shared sentence-transformer encoder (384-d) for embeddings; UMAP for dimensionality reduction; HDBSCAN for clustering; c-TF-IDF topic representations; noise topic retained. Key hyperparameters are, <ul style="list-style-type: none"> •UMAP: $n_{\text{neighbors}}=15$, $min_dist=0.0$, $n_{\text{components}}=5$. •HDBSCAN: $min_cluster_size=15$, $min_samples=10$.
Training data for topic model	Domain adaptation of the pre-trained BERTopic on the Food.com/Kaggle corpus ($\sim 180k$ recipes & interactions) prior to evaluation on the YouTube-8M food subset.
Evaluation protocol	Metrics: Top- k success ($k \in \{1, 2, 3, 5, 10\}$) and Coverage. Statistics: paired Wilcoxon signed-rank test with $\alpha=0.05$ comparing late vs. early fusion.
Data management & storage	All artefacts (keyframes, Rekognition JSON, ASR transcripts, manifests) stored in Amazon S3. A manifest (CSV) maps video ID \rightarrow scene IDs, keyframe timestamps, S3 URIs, and derived outputs for reproducibility.

mapping to ensure a common embedding space; cosine similarity scores are adjusted by topic/term intensities as defined in (4).

4 Experimental results and discussion

4.1 Test data

Two main types of datasets are used to evaluate our model. The first dataset given as a highly rich corpus applied to train the topic model; the second dataset comprises real videos that were applied to evaluate the topic analysis methodology. Basically, this model fine-tuned by the rich corpus was applied to real videos for identifying and analysing different topic.

4.1.1 Rich corpus training

The basic knowledge base obtained in this study will be important, considering that it will be treated as the training corpus for the training process of the pre-trained BERTopic model. This selected dataset, specific to the food domain, covers a diverse collection of articles and captures a wide range of semantic relationships and context. Its purpose is to provide the model with a complete understanding of the semantics, context, and nuances of the language. The pretrained BERTopic model undergoes training on this corpus, adapting its knowledge to the details of the food domain. This makes it context-sensitive and linguistically competent. For our specific objectives, we used a Kaggle dataset [67] comprising 180K+ cooking Recipes and Interactions extracted from Food.com. The adaptation of the pre-trained BERTopic model through training on this specific dataset highlights its adaptability and robustness within the culinary context. This provided a collection of 4349 distinct topics.

4.1.2 YouTube-8M videos for evaluation

To evaluate our model, a real set of videos with ground truth is essential. Among several public datasets, the YouTube-8M dataset was found to be more feasible, as it contains both video content and ground truth labels. Among the obstacles that were faced was selecting videos from broader categories, making it quite formidable to identify specific topics. To address this, we adopted a precise approach. We selected by filtering the categories with a particular emphasis on the food niche. This carefully curated YouTube-8M dataset takes stock of a diverse array of food-related videos and also includes a variety of visual and audio information. This approach enables us to validate both extremes. Audio transcripts and visually compelling videos, each having a strong presence, despite the potential noise in the other modality. In addition, it covers the middle ground, where both visual and audio modalities contribute equally to the representation of the video. This infused our analysis with deep insights into the food niche. As indicated in the methodology section, this line of videos was assessed in a detailed process where we first extract visual features in textual format for Visual model analysis and then audio features through audio transcripts for Audio modality analysis. The extracted visual words serve as input for the trained BERTopic model. This facilitates the identification of topics within the visuals. Moreover, the

audio transcripts serve as input for the same trained BERTopic model to identify topics within the audio, subsequently mapping them to standard context taxonomies.

4.2 Illustrating framework:A YouTube-8M video case study

An example of the industry taxonomies of representation for a YouTube-8M video with the proposed model is given in Table 2, which depicts the output probabilities of the topic modelling stage that, in turn, drives the final predictions of the contextual taxonomy representation. The comparison between the ground-truth taxonomy and the predictions shown in Fig. 5 verifies that it correctly mapped the first two predicted taxonomies, covering four associated ground truth taxonomies. Even the one that was missed is a close match when considering human evaluation. It is also important to note that our approach found other hidden themes in the YouTube-8M dataset that were not tracked in the ground truth, in addition to the mapped ones. Discovery of content related to children is one such example. Furthermore, the line graph shows how each of the modalities contributes to the final outcome.


4.3 Performance evaluation metrics

We applied the evaluation for a separately held dataset, so it was not an automatic process for an extensive dataset. Instead, we used a manual selection method to select 50 randomly chosen videos from the YouTube-8M dataset, all of which are in the speciality category of food videos. This subset aligns with the specific domain in which the framework was initially trained.

In evaluating the effectiveness of our framework for video taxonomy, we utilise a critical criterion, the degree of correspondence between the ground mapping. This is in relation to industry taxonomy and the predicted outcomes of taxonomy mappings generated by the framework.

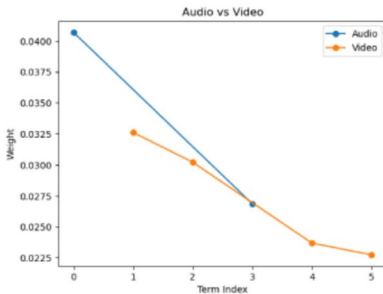
Our analysis is presented based on two key measures:

Table 2 Example analysis data summary

Example Analysis	
8M example video	
URL	https://www.youtube.com/watch?v=AMe56hxTMHw
Ground truth	'Food', 'Recipe', 'Cake', 'Cookie', 'Cupcake', 'Cake pop', 'Food', 'Recipe', 'Cake', 'Cookie', 'Cupcake', 'Cake pop'
Audio topics	[330, 1483, 899, 567, 79]
Audio topic weights	[0.68541145, 0.60544884, 0.59641486, 0.5956478, 0.59390974]
Video topics	[2701, 3929, 2752, 960, 2343]
Video topic weights	[0.4095165, 0.40553853, 0.40358847, 0.3900413, 0.35338795]
Video+audio topics	[330, 3239, 899, 567, 1367]
Video+audio topic weights	[0.60842794, 0.5249552, 0.5149646, 0.49994525, 0.49978724]

Prediction Taxonomies

	source	term	taxonomy_keyword	cosine_similarity	topic_weight	term_weight	final_weight
0	audio	treats	Desserts and Baking Food & Drink Desserts and ...	0.509367	0.596415	0.133958	0.040696
1	visual	foods	Food & Drink Food & Drink	0.620841	0.353388	0.148661	0.032616
2	visual	ideas	Model Toys Hobbies & Interests Model Toys	0.502574	0.403588	0.148953	0.030213
3	audio	candy	Gifts and Greetings Cards Shopping Gifts and G...	0.397023	0.685411	0.098664	0.026849
4	visual	everyday	Wellness Healthy Living Wellness	0.400660	0.390041	0.151518	0.023678
5	visual	rosengarten	Apprenticeships Careers Apprenticeships	0.295004	0.353388	0.218095	0.022737



Ground Truth Taxonomies

term	taxonomy_keyword	cosine_similarity
Food	Food & Drink Food & Drink	0.617123
Cake	Desserts and Baking Food & Drink Desserts and ...	0.534501
Cake pop	Desserts and Baking Food & Drink Desserts and ...	0.525742
Recipe	Cooking Food & Drink Cooking	0.502049
Cupcake	Desserts and Baking Food & Drink Desserts and ...	0.500328
Cookie	Desserts and Baking Food & Drink Desserts and ...	0.417892

Fig. 5 Ground truth vs. predicted taxonomy comparison in a sample and highlighting visual and audio modality contributions

- Evaluation of the top N predicted video taxonomy interpreted for sampled videos.
- Comparison of the proximity of the data from the predicted outcomes

These are key integral measures used to evaluate the performance of the framework in the target domain. It puts more emphasis on the match between predicted and standard industry taxonomy mappings to give a more subtle and credible evaluation.

4.3.1 Evaluation of the top N predicted video taxonomy interpreted for sampled videos

This offers a comprehensive analysis of the performance of the model at different confidence levels, including Top 1, Top 2, etc. The metric is derived from the first correct prediction at which confidence level within the top n predicted labels. Generates valuable information on the reliability of the model at different prediction confidence levels.

The evaluation score is calculated as follows:

Let $P = \{L_1, L_2, \dots, L_m\}$ represent the set of taxonomies of ground truth labels for a video, and let $T_i = \{P_1, P_2, \dots, P_n\}$ denote the set of taxonomies of predicted labels in the top n predictions.

Define an indicator function δ_k for each top k as:

$$\delta_k = \begin{cases} 1 & \text{if } P_k \cap T_i \neq \emptyset \text{ (i.e., there is an intersection)} \\ 0 & \text{otherwise} \end{cases}$$

The “Overall Evaluation” is then calculated as:

$$\text{Overall Evaluation} = \sum_{k=1}^n \delta_k$$

This metric counts the number of true tops from 1 to n , where any single correct prediction is counted, to give full comprehension of the model’s performance under different confidence levels. Given that YouTube-8M videos often include multiple topic labels, usually not exceeding a certain number (E.g., 10), we focused our experiments on determining the top N -predicted values with the highest success scores. We systematically varied and tested the value of N at 1, 2, 3, 5, and 10, respectively. The experiments were carried out on a test set with more than 50 YouTube-8M videos. The results are presented in Table 3. The tests give insights into how well our model can perform if we consider various levels of prediction confidence.

The results suggest that the late-fusion model demonstrated improvement. Furthermore, this model consistently exhibited higher percentages for the 50 selected videos in the experiment, achieving values of 90.7% and 92.7% for predictions within the top 5 and top 10, respectively.

4.3.2 Compare the proximity of the data from the predicted data

Although the Jaccard similarity coefficient method is commonly employed to assess the similarity between data points in the context of data, its effectiveness is hindered by specific characteristics in our dataset. The presence of videos within the same niche domain results in a limited number of unique video topic labels after taxonomy association, attributed to their convergence into similar vector embeddings. Consequently, a few mismatches can disproportionately impact the evaluation, leading to an inaccurate representation of the model performance. To address this, we introduced the “Coverage” metric, which calculates the number of predicted taxonomies versus the ground truth-associated taxonomies. This alternative metric provides a more contextually relevant evaluation of the framework performance in our specific dataset conditions.

Let N_V be the total number of videos, N_{PT} be the total number of predicted taxonomies, and N_{GT} be the total number of taxonomies associated with ground truth. The Average Coverage metric is calculated as follows:

$$\text{Average Coverage} = \frac{1}{N_V} \sum_{i=1}^{N_V} \left(\frac{N_{PT_i}}{N_{GT_i}} \right) \times 100$$

Table 3 Success Rate for Initial Correct Predictions at Top Confidence Levels in Predicted Labels

Fusion	Top1	Top 2	Top 3	Top 5	Top 10
Late	52.8%	73.6%	81.1%	90.7%	92.7%
Early	46.2%	69.8%	77.4%	81.1%	90.6%

This metric expresses the percentage of predicted taxonomies compared to the ground truth-associated taxonomies, offering insights into the framework’s performance in capturing the relevant taxonomies within the tested dataset.

The results presented in Table 4 summarise the results of the early and late fusion strategies.

We should point out that the BERTopic model has been fine-tuned on a domain-specific corpus of text related to food. Due to that, the topic inference leans towards food-oriented topics. Although transfer learning, as well as semantic generalization, may still capture some of the other thematic elements, their coverage remains inherently constrained because of the limited domain of the corpus used for training. This partial coverage contributed to a lower overall taxonomy coverage score (41.2% and 36.5%). To overcome such a limitation, we aim to make the training corpus cover a broad set of industry-specific domains. It will enable the model to capture more diversified themes and achieve enhanced coverage as well as contextually aligned results for more extended categories of videos (Table 5).

4.3.3 Statistical validation of results

To rigorously assess our results, we applied the standard hypothesis test between our *late-fusion* and *early-fusion* techniques. The following hypotheses were formulated:

- H_0 : Late and early fusion performance is not significantly different.
- H_1 : Late fusion vastly overwhelms early fusion.

We applied paired Wilcoxon signed-rank tests (the non-parametric equivalent of the *t*-test, suitable for our sample size $n = 50$) on the following metrics of evaluation:

- **Top-5 Success Rate** (primary metric)
- **Coverage Metric**

Results

Interpretation With p -values < 0.05 for both metrics, we reject H_0 and conclude that *late-fusion* significantly outperforms *early-fusion*. The moderate effect sizes ($r > 0.3$) confirm practical significance beyond statistical significance.

Table 4 Taxonomy Prediction Coverage (%)

Fusion	Coverage
Late	41.2%
Early	36.5%

Table 5 Comparison of Late Fusion and Early Fusion techniques using paired Wilcoxon signed-rank tests

Metric	Late Fusion	Early Fusion	P-value	Effect Size (r)
Top-5 Success	90.7%	81.1%	0.003	0.42
Coverage	41.2%	36.5%	0.018	0.31

5 Discussion

5.1 Multi-Dimensional benchmarking with state-of-the-art methods

As part of verifying further the performance of our framework, we performed extensive comparison with the latest contextual advertising state-of-the-art techniques. Given the diversity of evaluation metrics used in recent research on video-based advertising, we adopt a multi-dimensional benchmarking methodology that compares our approach with prior studies in terms of methodology, capability, and performance characteristics. Table 6 gives a comprehensive comparison of our model with two of the recently introduced state-of-the-art methods, SemanticAd (2024) [68] and “What Modality Matters” (2023) [69]. It includes comparison across many of the key dimensions like core objectives, input modalities, key contributions, evaluation approaches, and other important points.

This multi-faceted comparison yields multiple insights with reference to the positioning of our approach with respect to prior work. First, while prior work specialized within specific dimensions of the contextual ad task like modality inference or time partitioning, our approach holistically assumes the broader task of video content to taxonomies corresponding within an industry-agnostic manner. Second, the use of BERT embeddings forms a unique strength for handling large-scale, dynamic content settings, for effective indexability and information lookup along with domain adaption with minimal retraining. Third,

Table 6 Multi-dimensional comparison highlighting distinctive aspects of our framework

Dimension	SemanticAd (2024)	What Modality Matters (2023)	Our Framework
Core Technical Approach	Temporal segmentation of video into story units	Modality importance analysis for relevance scoring	Topic modeling with industry taxonomy mapping
Semantic Understanding	Scene-level content analysis	Multi-modal feature extraction	Deep topic semantics using BERTopic with contextual fusion
Industry Readiness	Ad insertion point detection	Relevance scoring for ad matching	Standardized taxonomy output compatible with IAB standards
Technology Used	Shot-boundary detection, temporally constrained clustering, multimodal fusion	Pretrained CNNs (TSN, YOLOv5, PlacesCNN, EmotionNet, CNN14) with similarity matching	BERTopic topic modelling and inferencing, vector-space embedding similarity, IAB taxonomy mapping
Evaluation Approach	Custom dataset (50 TV news videos)	Movie content + 14 ads with expert study	YouTube-8M with ground truth taxonomy labels
Performance Metrics	F1-score (92%+)	Relevance Score (S) (0.337)	Top-5 Success Rate (90.7%)

by leveraging topic modelling (BERTopic), our approach is able to rapidly respond to new themes evolving within the video content, enabling timely ad content alignment with popular topics without deep re-engineering of the targeting workflow.

5.2 Addressing class imbalance and database bias

Class imbalance and dataset bias are two significant problems of deep learning from large datasets like YouTube-8M. As observes [70], “bigger datasets are not always better” in the instance of data bias or class imbalance. Three key mitigation measures we adopted were: First, Targeted Dataset Selection: We used a food-domain subset rather than the entire YouTube-8M dataset to reduce imbalance through the identification of a specific domain. Our test used the application of 50 food videos from the same specialty category. Second, Model Design: BERTopic computes optimal numbers of topics automatically and uses “noise topics” to prevent misalignment of noisy documents—most useful for imbalanced data. Third, Overall Evaluation: We applied a variety of metrics (Top-N success rates, Coverage) instead of accuracy only, and incorporated human evaluations as well as automated evaluations to identify bias problems. Despite these measures, limitations remain. Our evaluation only took into account videos from the food category and thus possibly misses the YouTube-8M richness. Also, we did not systematically look into dataset biases (e.g., overrepresentation of particular video classes). These issues are worthy of attention in the future, as claimed [70].

5.3 Multimodal fusion performance and framework limitations

Our statistical validation (Section 4.3.3) provides rigorous evidence that the *late-fusion* approach clearly outperforms the *early-fusion* ($p < 0.01$). This overrules potential scepticism regarding the reliability of results posed by the properties of datasets [70]. The test of hypothesis supports our claims regarding the effectiveness of multimodal fusion while embracing natural limitations on the video datasets analysis. Our novel method shows that using both visual and audio information improves the precision and detail of content classification and mapping to content taxonomy. We observed that our framework could identify hidden themes not tracked in the coverage due to the absence of specific labels in the ground truth for taxonomy alignments. This deficiency had a negative impact on coverage. Our research showed that audio transcripts had a greater impact on the probability weights for audio topics than visual topics, resulting in a higher contribution of audio topic terms to the final result. However, visual themes were more prominent when there was audio noise. Additionally, the early fusion model worked well when there was a lot of information exchange and coherence between visual and audio features. Current studies agree that the use of a multimodal approach improves content recognition and classification outcomes.

One of the key limitations of this approach will be to consider all the returned image objects by AWS Rekognition as visual words. This might introduce background noise, as objects which are detected do not always contribute meaning to the inherent theme of the video. It might also make visual topic modelling more discriminative by adding scene knowledge or scoring of object significance. Additionally, this constraint is a result of trying to utilize the full potential of BERTopic particularly on the image side, for which we do not yet have the capacity to understand the semantic significance of identified objects. This has been noted as a future work high priority, where we will incorporate semantic knowledge of

visual features towards enhancing topic coverage as well as relevance. Another area where enhancement can be focused is, in topic representation. Existing methods overlay the topic by counting topic words that may not give a detailed and in-depth comprehension of themes. In this regard, the process may be further refined if coupled with topic representations, thus giving a detailed and thorough comprehension of the video content in the process. An emerging direction worth exploring is the integration of large language models (LLMs), which offer the potential to produce semantically rich and interpretable topic representations that go beyond shallow word frequency counts. All these challenges can be overcome to scale up the accuracy and efficiency of thematic ideation and content mapping for ad discovery to new heights, enabling more precise and lean advertising.

5.4 Positioning against synthetic media

Increasing popularity of synthetic media (deepfakes, etc.) presents threats to contextual ad systems: adversarially crafted visual or audio streams could induce spurious topics and thus unsuitable taxonomy labels with brand-safety implications. Our system counters those threats with cross-modal coherence (late fusion requires consistent evidence across modalities), topic-level semantics (neural topic modeling prefers thematic coherence to surface features), and taxonomy-level verification (canonicalized label alignment highlights inconsistencies typical of synthetic materials). In practice, these features assist auditing of contents and anomaly detection—i.e., flags videos where audio–visual semantics are at odds or where taxonomy confidence profiles areypical of corpus. While not a special-purpose forgery detector, the system is an adequate accompaniment to authentication checks; adding explicit detection of deepfakes is a natural extension with future promise.

5.5 Limitations and future work

Our framework, while effective, has several limitations that motivate future work. First, domain generalization is constrained: the BERTopic adaptation was tuned on food-domain data and evaluated on a food-focused subset, suggesting the need for multi-domain pretraining and zero/few-shot tests. Second, noisy visual labels from generic detectors can bias topics; integrating scene understanding, saliency scoring, or visual captioning could yield richer descriptors. Third, ASR errors under adverse acoustics can skew audio topics; robustness can improve with confidence-aware fusion and speech enhancement. Fourth, outputs are largely video-level; adding temporal segmentation would enable time-stamped taxonomy labels for fine-grained ad insertion and brand safety. Fifth, fusion is a simple convex weighting rather than a learned, content-adaptive gate; lightweight multimodal transformers or gating modules could learn better fusion. Sixth, confidence scores are not calibrated; post-hoc calibration and abstention policies would improve reliability. Finally, evaluation scale and deployment aspects (privacy/federation, long-tailed labels, robustness to synthetic media) warrant larger multi-domain studies, federated variants, bias mitigation, and integration of deepfake detection.

6 Conclusion

The proposed novel framework highlights the timely issue of contextual advertising, given the emerging regulations to track user behaviour and move toward the cookieless environment. By effectively applying contextual video analysis with industry context taxonomy, our proposed novel framework functions effectively for effective advertising within advanced advertising ecosystems. Combination of video and audio thematic analysis helps bridge the serious semantic gap in a bid to make the translation of video content into more enhanced persuasive contextual participation. In this work, we apply neural components for an effective and efficient association of video content with context-based taxonomy. Our experimental validation on the YouTube-8M dataset validates the viability of the framework. Top 5 prediction success achieved 90.7% and top 10 prediction success achieved 92.7% by applying the late-fusion scheme, outperforming early-fusion schemes (p -values < 0.05).

In our future work, we will elaborate on the use of large language models to achieve topic explainability within our framework, this currently represents an open challenge. Sentiment analysis within this system, particularly for negative sentiment in videos, is another area for future work that we are considering. This considers sentiment as implicitly subjective and discusses it in the context of the advertisers' messages. While our system provides good contextual analysis for advertisement deployment, we also value the dynamic challenges posed by the generation of fake content. Our future work will be aimed toward the integration of certain deepfake detection features and authenticity checking processes in order to maintain the ongoing integrity of contextual advertisement systems.

Author Contributions WDS conducted the entire study, including conceptualization, data collection, analysis, and manuscript writing. AF provided supervision, guidance, and critical review of the manuscript. All authors read and approved the final manuscript.

Funding The authors received no funding for this work.

Data Availability Dataset employed in this research paper constitutes a filtered portion of the publicly available dataset, YouTube-8M dataset (<https://research.google.com/youtube8m/>). We picked 50 videos from the category "Food" to be included in our analysis. The original dataset can be accessed from the link provided. The filtered set used in this research paper can be obtained from the corresponding author upon reasonable request.

Declarations

Competing interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Chalaby JK (2024) The streaming industry and the platform economy: An analysis. *Media Cult Soc* 46(3):552–571
2. Miller KM, Skiera B (2024) Economic consequences of online tracking restrictions: Evidence from cookies. *Int J Res Mark* 41(2):241–264
3. Cooper DA, Yalcin T, Nistor C, Macrini M, Pehlivan E (2023) Privacy considerations for online advertising: A stakeholder's perspective to programmatic advertising. *J Consum Mark* 40(2):235–247
4. Rohrer C, Boyd J (2004) In: CHI'04 extended abstracts on human factors in computing systems, pp 1085–1086
5. Li H, Edwards SM, Lee J-H (2002) Measuring the intrusiveness of advertisements: Scale development and validation. *J Advert* 31(2):37–47
6. Singh M, Lamba R (2020) Proposing contextually relevant advertisements for online videos. In: Machine learning and metaheuristics algorithms, and applications: first symposium, SoMMA 2019, Trivandrum, India, December 18–21, 2019, Revised Selected Papers 1, pp 218–224. Springer
7. Mei T, Yang L, Hua X-S, Wei H, Li S (2007) Videosense: A contextual video advertising system. In: Proceedings of the 15th ACM international conference on multimedia, pp 463–464
8. Xiang C, Nguyen TV, Kankanhalli M (2015) Salad: A multimodal approach for contextual video advertising. In: 2015 IEEE International symposium on multimedia (ISM), pp 211–216. IEEE
9. Okada K, Moura ES, Cristo M, Fernandes D, Gonçalves MA, Bertl K (2012) Advertisement selection for online videos. In: Proceedings of the 18th Brazilian symposium on multimedia and the web, pp 367–374 (2012)
10. Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: A survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell* 41(2):423–443
11. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4694–4702
12. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision, pp 20–36. Springer
13. Wu Q, Zhu A, Cui R, Wang T, Hu F, Bao Y, Snoussi H (2021) Pose-guided inflated 3d convnet for action recognition in videos. *Signal Process Image Commun* 91:116098
14. Tang H, Ding L, Wu S, Ren B, Sebe N, Rota P (2023) Deep unsupervised key frame extraction for efficient video classification. *ACM Trans Multimed Comput Commun Appl* 19(3):1–17
15. Drossos K, Adavanne S, Virtanen T (2017) Automated audio captioning with recurrent neural networks. In: 2017 IEEE Workshop on applications of signal processing to audio and acoustics (WASPAA), pp 374–378. IEEE
16. Wu M, Dinkel H, Yu K (2019) Audio caption: Listen and tell. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 830–834. IEEE
17. Yu D, Deng L (2016) Automatic Speech Recognition vol 1. Springer
18. Blei DM (2012) Probabilistic topic models. *Commun ACM* 55(4):77–84
19. Interactive Advertising Bureau (2023) Interactive advertising bureau taxonomies. <https://github.com/InteractiveAdvertisingBureau/Taxonomies>. Accessed 9 Dec 2023
20. Wood J, Tan P, Wang W, Arnold C (2017) Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In: 2017 IEEE 33rd International conference on data engineering (ICDE), pp 411–422. <https://doi.org/10.1109/ICDE.2017.99>
21. Lau JH, Newman D, Karimi S, Baldwin T (2010) Best topic word selection for topic labelling. In: *Coling 2010: Posters*, pp 605–613
22. Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan S (2016) Youtube-8m: A large-scale video classification benchmark. [arXiv:1609.08675](https://arxiv.org/abs/1609.08675)
23. Zhang W, Rong Y, Wang J, Zhu T, Wang X (2016) Feedback control of real-time display advertising. In: Proceedings of the ninth ACM international conference on web search and data mining, pp 407–416
24. Zhang H, Ji Y, Huang W, Liu L (2019) Sitcom-star-based clothing retrieval for video advertising: a deep learning framework. *Neural Comput Appl* 31:7361–7380
25. Wang G, Zhuo L, Li J, Ren D, Zhang J (2018) An efficient method of content-targeted online video advertising. *J Vis Commun Image Represent* 50:40–48
26. Luo C, Peng Y, Zhu T, Li L (2019) An optimization framework of video advertising: using deep learning algorithm based on global image information. *Clust Comput* 22(Suppl 4):8939–8951
27. Song X, Xu B, Jiang Y-G (2020) Predicting content similarity via multimodal modeling for video-in-video advertising. *IEEE Trans Circuits Syst Video Technol* 31(2):569–581

28. Tapu R, Mocanu B, Zaharia T (2020) Deep-ad: a multimodal temporal video segmentation framework for online video advertising. *IEEE Access* 8:99582–99597
29. Mocanu B, Tapu R (2024) Semanticad: A multimodal contextual advertisement framework for online video streaming platforms. *IEEE Access* 12:63142–63155. <https://doi.org/10.1109/ACCESS.2024.3395922>
30. Joa CY, Kim K, Ha L (2018) What makes people watch online in-stream video advertisements? *J Interact Advert* 18(1):1–14
31. Xue J, Eguchi K (2018) Sequential bayesian nonparametric multimodal topic models for video data analysis. *IEICE Trans Inf Syst* 101(4):1079–1087
32. Thies J, Stappen L, Hagerer G, Schuller BW, Groh G (2021) Graphmt: Unsupervised graph-based topic modeling from video transcripts. In: 2021 IEEE seventh international conference on multimedia big data (BigMM), pp 1–8. <https://doi.org/10.1109/BigMM52142.2021.00009>
33. Rafiq G, Rafiq M, Choi GS (2023) Video description: A comprehensive survey of deep learning approaches. *Artif Intell Rev* 56(11):13293–13372
34. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738
35. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning, pp 1597–1607. PmLR
36. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp 8748–8763. PmLR
37. Yang H, Huang D, Wen B, Wu J, Yao H, Jiang Y, Zhu X, Yuan Z (2022) Self-supervised video representation learning with motion-aware masked autoencoders. [arXiv:2210.04154](https://arxiv.org/abs/2210.04154)
38. Huang B, Zhao Z, Zhang G, Qiao Y, Wang L (2023) Mgm: Motion guided masking for video masked autoencoding. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 13493–13504
39. Hernandez J, Villegas R, Ordonez V (2024) Vic-mae: Self-supervised representation learning from images and video with contrastive masked autoencoders. In: European conference on computer vision, pp 444–463. Springer
40. Sun C, Myers A, Vondrick C, Murphy K, Schmid C (2019) Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7464–7473
41. Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding? In: *Icml*, vol 2, p 4
42. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6836–6846
43. Chen J, Zhang H, Gong M, Gao Z (2024) Collaborative compensative transformer network for salient object detection. *Pattern Recogn* 154:110600
44. Wang Y, Li J, Zhang H, Zhang J, Wan F, Qiu A, Gao Z (2025) Fedmdd: Multi-deliberation based calibration for federated long-tailed learning. *Knowl-Based Syst* 113741
45. Sharma V (2022) Object detection and recognition using amazon rekognition with boto3. In: 2022 6th international conference on trends in electronics and informatics (ICOEI), pp 727–732. <https://doi.org/10.1109/ICOEI53556.2022.9776884>
46. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I (2023) Robust speech recognition via large-scale weak supervision. In: International conference on machine learning, pp 28492–28518. PMLR
47. Jiao T, Guo C, Feng X, Chen Y, Song J (2024) A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Comput Mat Continua* 80(1):1–35. <https://doi.org/10.32604/cmc.2024.053204>
48. Alghamdi R, Alfalqi K (2015) A survey of topic modeling in text mining. *Int J Adv Comput Sci Appl (IJACSA)* 6(1)
49. Sivic, Zisserman (2003) Video google: A text retrieval approach to object matching in videos. In: Proceedings ninth IEEE international conference on computer vision, pp 1470–1477. IEEE
50. Zhou L, Zhou Z, Hu D (2013) Scene classification using a multi-resolution bag-of-features model. *Pattern Recogn* 46(1):424–433
51. Silva FB, Werneck RdO, Goldenstein S, Tabbone S, Torres RdS (2018) Graph-based bag-of-words for classification. *Patt Recogn* 74:266–285
52. Barde BV, Bainwad AM (2017) An overview of topic modeling methods and tools. In: 2017 International conference on intelligent computing and control systems (ICICCS), pp 745–750. IEEE
53. Blei D, Carin L, Dunson D (2010) Probabilistic topic models. *IEEE Signal Process Mag* 27(6):55–65

54. Wang W, Barnaghi PM, Bargiela A (2009) Probabilistic topic models for learning terminological ontologies. *IEEE Trans Knowl Data Eng* 22(7):1028–1040
55. Lau JH, Newman D, Baldwin T (2014) Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th conference of the european chapter of the association for computational linguistics, pp 530–539
56. Stevens K, Kegelmeyer P, Andrzejewski D, Buttler D (2012) Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp 952–961
57. Grootendorst M (2022) Bertopic: Neural topic modeling with a class-based tf-idf procedure. [arXiv:2203.05794](https://arxiv.org/abs/2203.05794)
58. Cao Z, Li S, Liu Y, Li W, Ji H (2015) A novel neural topic model and its supervised extension. In: Proceedings of the AAAI conference on artificial intelligence, vol 29
59. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (Long and Short Papers), pp 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. <https://doi.org/10.18653/v1/N19-1423>. <https://aclanthology.org/N19-1423/>
60. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3(Jan):993–1022
61. Egger R, Yu J (2022) A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Front Sociol* 7:886498
62. Rahimi H, Naacke H, Constantin C, Amann B (2024) Antm: aligned neural topic models for exploring evolving topics. In: Transactions on large-scale data-and knowledge-centered systems LVI: special issue on data management-principles, technologies, and applications, pp 76–97. Springer, ???
63. McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. [arXiv:1802.03426](https://arxiv.org/abs/1802.03426)
64. Van Gemert JC, Veenman CJ, Smeulders AW, Geusebroek J-M (2009) Visual word ambiguity. *IEEE Trans Pattern Anal Mach Intell* 32(7):1271–1283
65. Van De Sande K, Gevers T, Snoek C (2009) Evaluating color descriptors for object and scene recognition. *IEEE Trans Pattern Anal Mach Intell* 32(9):1582–1596
66. Hall D, Jurafsky D, Manning CD (2008) Studying the history of ideas using topic models. In: Proceedings of the 2008 conference on empirical methods in natural language processing, pp 363–371
67. Li S (2019) Food.com recipes and user interactions. <https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions>
68. Mocanu B, Tapu R (2024) Semanticad: A multimodal contextual advertisement framework for online video streaming platforms. *IEEE access* 12:63142–63155
69. Chong OK, Goh H-N, See J (2023) What modality matters? exploiting highly relevant features for video advertisement insertion. In: 2023 IEEE international conference on image processing (ICIP), pp 3344–3348. IEEE
70. Roccetti M, Delnevo G, Casini L, Cappiello G (2019) Is bigger always better? a controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. *J Big Data* 6(1):1–23

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.