Scientific Data

https://doi.org/10.1038/s41597-026-07364-3

Article in Press

A dataset of four Decades of Italian TV Commercials: visual, audio, and linguistic feature descriptors

Received: 31 October 2025

Accepted: 28 April 2026

Published online: 08 July 2026

Cite this article as: Fadda, D., Bellante, G., Rinzivillo, S. A dataset of four Decades of Italian TV Commercials: visual, audio, and linguistic feature descriptors. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-07364-3

Daniele Fadda, Giorgio Bellante & Salvatore Rinzivillo

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

ARTICLE IN PRESS

A dataset of four Decades of Italian TV Commercials: visual, audio, and linguistic feature descriptors

Daniele Fadda*[1], Giorgio Bellante[2], and Salvatore Rinzivillo[1]

[1]Institute of Information Science and Technologies (ISTI), National Research Council, Pisa, Italy
[2]University of Pisa, Italy
*corresponding author: Daniele Fadda (daniele.fadda@isti.cnr.it)

ABSTRACT

We present a feature-level dataset of Italian television commercials spanning over four decades (1980–2024) and including more than 10,000 adverts. Unlike existing corpora, which are often limited in scope or duration, this resource captures the longitudinal evolution of a national advertising ecosystem through structured descriptors rather than raw media. Each commercial is enriched with metadata (title, brand, product category, year...) and segmented into scenes with representative frames from which we extract standardised colour palettes. We analyse audio tracks to extract spectral, chromatic, and rhythmic features, while spoken content is processed into linguistic annotations and psychometric markers. This integrated collection of visual, auditory, and textual descriptors enables cross-modal and temporal studies of advertising strategies and cultural trends. By releasing the dataset openly together with the complete preprocessing code, we ensure transparency and reproducibility, while also providing a methodology that can be readily applied to expand this collection or to build similar resources in other cultural and linguistic contexts, thereby supporting research in communication studies, multimedia analytics, and computational social science.

ARTICLE IN

Background & Summary

Television commercials are cultural artefacts that mirror social trends, marketing strategies, and creative practices over time. However, systematic analyses of advertising content across long periods have been limited by the scarcity of large, structured datasets. Several studies have examined the societal impact of commercials on consumer behaviour, for example investigating the relationship between televised advertising and child nutrition[1,2]. These works highlight the potential influence of advertising on habits, yet they mainly rely on broadcast metadata (e.g., airing frequency, channel, or product category) rather than on the internal composition of the ads, largely due to the absence of large-scale annotated corpora.

Recent initiatives have begun to address this limitation by compiling advertising datasets, such as large repositories of Japanese television commercials enriched with audience impression ratings[3], or multilingual collections of advertisement images organised into hierarchical taxonomies[4]. These resources underscore the growing potential of data-driven research in advertising studies and have stimulated new methodological contributions that stress the need for systematic computational content analysis integrating text, image, audio, and video to examine not only the conveyed message but also its executional aspects[5]. Nonetheless, most existing datasets still overlook fundamental expressive components of advertising language, such as rhythm and pacing, chromatic features, and the role of non-verbal audio elements including music and sound effects.

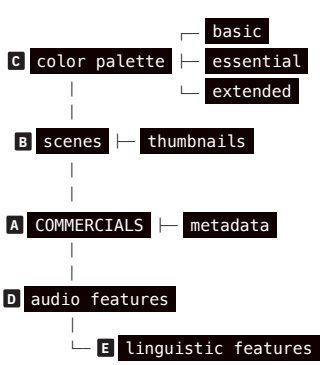
The dataset presented in this article addresses these limitations by focussing on Italian television commercials produced between 1980 (the year that marks the advent of commercial television in Italy) and 2024. It comprises more than 10,300 commercials, making it one of the most extensive national collections of advertisements enriched with multimodal descriptors. The materials were sourced from publicly available online archives and digital platforms, providing both historical and contemporary coverage as well as a broad representation of product categories and stylistic domains. Metadata such as title, brand, year of airing, and product category were manually curated and standardised to ensure interoperability. Spanning over four decades, the dataset enables systematic longitudinal analysis of trends in the evolution of advertising communication.

A distinctive feature of this dataset is the multimodal representation of each commercial. While the video files themselves are not distributed, we provide structured information derived from their visual, audio, and linguistic components. Each commercial is automatically segmented into scenes, providing fine-grained data on editing pace and narrative structure. For each scene, a representative frame is extracted and analysed to generate colour palettes that describe the chromatic composition of the ad. On the audio side, low-level acoustic descriptors quantify rhythmic, spectral, and timbral properties, while spoken content is transcribed and annotated linguistically and psycholinguis-

Files Structure

```
.
├── general A
│   ├── commercials.csv
│   ├── nice_classes.csv
│   ├── product_types.csv
│   └── scenes.csv
├── thumbnails.zip B
├── colors C
│   ├── basic_palette.csv
│   ├── basic_palette_idfs.csv
│   ├── commercial_palettes.csv
│   ├── essential_palette.csv
│   ├── essential_palette_idfs.csv
│   ├── extended_palette.csv
│   ├── extended_palette_idfs.csv
│   └── reference_palette_hierarchy.csv
├── audio D
│   ├── audio_feature_names.csv
│   ├── features
│   │   ├── chroma_cens_files.zip
│   │   ├── chroma_cqt_files.zip
│   │   ├── chroma_stft_files.zip
│   │   ├── librosa_static_tempo_default...
│   │   ├── librosa_static_tempo_lognorm...
│   │   ├── librosa_static_tempo_uniform...
│   │   ├── melspectrogram_files.z01
│   │   ├── melspectrogram_files.z02
│   │   ├── melspectrogram_files.zip
│   │   ├── mfcc_files.zip
│   │   ├── poly_features_files.zip
│   │   ├── rms_files.zip
│   │   ├── spectral_bandwidth_files.zip
│   │   ├── spectral_centroid_files.zip
│   │   ├── spectral_contrast_files.zip
│   │   ├── spectral_flatness_files.zip
│   │   ├── spectral_rolloff_files.zip
│   │   ├── tempocnn_static_tempo_files.zip
│   │   ├── tempogram_files.zip
│   │   ├── tonnetz_files.zip
│   │   └── zero_crossing_rate_files.zip
│   └── speech_class_confidence_scores.csv
└── text E
    ├── lemmas.csv
    └── liwc_analysis.csv
```

Dataset overview



Dataset in numbers

from 1980 to 2024	106k scenes detected	801 reference colors	197k audio feature files	97 psychological features
10,350 commercials	106k thumbnails	3.4M colors extracted	19 audio features	391k lemmas

Commercial example: Giochi Preziosi (Winx Club, 2015)

A Metadata

TITLE
Winx Club - Super Pasqualone Giochi Preziosi 2015! (SPOT TV)

AIRING DATE
2015

DURATION
15″

ASPECT RATIO
16:9

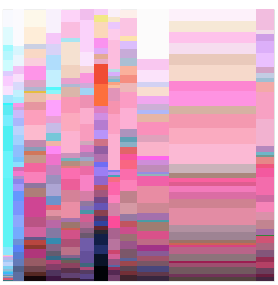
NICE CLASS
28. Games, toys and playthings; video game apparatus; gymnastic and sporting articles; decorations for Christmas trees

PRODUCT TYPE
Toys, games, and playthings

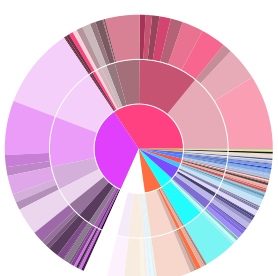
SOURCE
Winx Club Italia (YouTube)

B Scene thumbnails

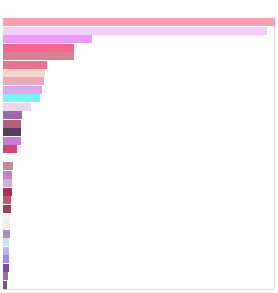
C Color analysis



Original color palette extracted from scene frames, sorted by lightness

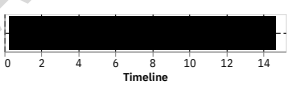


Commercial-level palette mapped to the reference color system, sorted by frequency



TF-IDF ranking of colors within the commercial, relative to the essential palette hierarchy

D Audio features



Speech segments detected throughout the commercial

137 BPM

The tempo, representing the ad's rhythmic pace

E Linguistic features

LEMMA	UPOS
Winx	PROPN
Pasqualone	PROPN
Show	PROPN
Music	PROPN
microfono	NOUN
trovare	VERB
Bloom	PROPN
cantare	VERB
Super	PROPN
magia	NOUN

Top 10 lemmas and part-of-speech tags extracted from transcription

ARTICLE IN PRESS

Figure 1. The figure provides an overview of the dataset structure, contents, and analytical dimensions. The left column (Files Structure and Dataset overview) summarises the corpus composition and its organisation into visual, auditory, and textual components. Letters **A–E** identify the five core analytical dimensions consistently referenced across the figure: **A)** metadata; **B)** scene thumbnails; **C)** color analysis; **D)** audio features; and **E)** linguistic features. The example commercial (Winx Club – Super Pasqualone Giochi Preziosi 2015) illustrates how some of these dimensions are integrated: structured metadata **(A)**, automatically segmented scenes **(B)**, hierarchical color palettes derived from frames **(C)**, detected speech and tempo information **(D)**, and linguistic descriptors extracted from transcriptions **(E)**.

ARTICLE IN PRESS

tically. This combination of visual, auditory, and textual descriptors enables cross-modal studies of communication strategies, narrative patterns, and aesthetic trends. A comprehensive overview of the dataset structure and some analytical dimensions is provided in Figure 1.

The broader goal of this work is to support research in the social sciences and communication studies that examines what commercials contain, rather than merely when or where they aired. By structuring the data in a machine-readable format and providing complete processing code, the dataset lowers the technical barrier for applying computational methods to advertising analysis. Potential works include examining gender and age representation, studying the evolution of stylistic and creative techniques (for example dominant colours, lexical patterns, or tempo), and training machine learning models to predict visual or linguistic characteristics of commercials. Unlike traditional content analyses based on manual annotation of small samples, this dataset enables scalable and reproducible research across thousands of ads.

The proposed resource is projected to hold considerable significance for advertising and communication research, as well as for the computer vision, audio analysis, and natural language processing communities, by providing a robust and standardised benchmark for multimodal content understanding. The dataset is publicly available via Zenodo[6], together with the complete preprocessing code, ensuring transparency, reproducibility, and reusability.

By sharing data and methods, we offer a framework to build similar datasets in other contexts, languages and regions, supporting research on advertising as a reflection of social change.

Methods

This section outlines the procedures adopted to create and enrich the dataset of Italian television commercials. The methodological workflow covers data collection, preprocessing, feature extraction and structuring of visual, auditory, and textual information in a format suitable for computational analysis and reuse.

Data Collection and Preprocessing

Commercial videos were collected from multiple online sources to achieve broad coverage across years and product sectors. The primary sources were an Italian digital archive called Spot 80[7] (an independent website curating vintage commercials) and more than 500 YouTube channels dedicated to TV commercials. In particular, the Spot 80 archive contributed 1,734 commercials (primarily covering 1977–1999), a YouTube account “mDeplo” provided over 6,700 commercials (mostly grouped in compilation videos), and another channel “Engage.it” contributed 199 commercials from the 2010s. In addition, more than 1,600 ads were gathered from various other YouTube users, collectively spanning hundreds of different brands and campaigns. In total, 543 distinct YouTube users were analysed, of which 352 had uploaded only a single video. For temporary video download during the analysis phase, the YouTube Premium feature was used. The collected commercials vary in length from a few seconds up to about one minute, with most being standard 15-second or 30-second spots. All videos were downloaded in their highest available quality and converted to a consistent format for processing.

Regarding the selection criteria and corpus representativeness, the dataset was structured to systematically cover the entire evolutionary trajectory of commercial television in Italy from 1980 to 2024. Content selection was not arbitrary but grounded in the availability of carefully curated historical archives. The compilation process followed a hybrid opportunistic-targeted strategy: initially drawing from large, well-documented collections and subsequently expanding through the focused integration of videos retrieved from YouTube. This expansion utilized the specific query `"spot tv"` –a pseudo-Anglicism used in Italy to denote commercials– prioritizing results that facilitated the semi-automatic extraction of essential metadata such as brand, airing year, and product name. This approach ensured a high-quality enrichment of the `commercials` metadata table and allowed us to fill previously underrepresented time periods, resulting in a consistent longitudinal distribution across the advertising ecosystem.

Following this targeted selection, we compiled an initial metadata table recording for each video its original filename or YouTube ID, the title (as provided by the source), the duration, the frame rate, and the source identifier (e.g., “Spot80” or the YouTube channel name). This raw dataset was then cleaned and harmonised through several steps.

First, we removed duplicate commercials by checking for identical titles and overlapping metadata; we retained any repeated entries (for example, the same ad uploaded by multiple sources) only once. Furthermore, we manually annotated all available descriptive attributes (such as year, brand, and product category), with automation techniques—primarily based on regular expressions—applied wherever possible to support the process.

Second, we inspected videos for formatting issues: some recordings displayed black bars on sides (pillarbox effect) or along the top and bottom edges (letterbox effect). We performed a cropping operation to remove these bars and restore the correct aspect ratio when needed. As a result, the dataset includes aspect ratio information (primarily 4:3 for older TV footage and 16:9 for newer content) with all videos in true full-frame format (for an example Figures 2).

ARTICLE IN PRESS

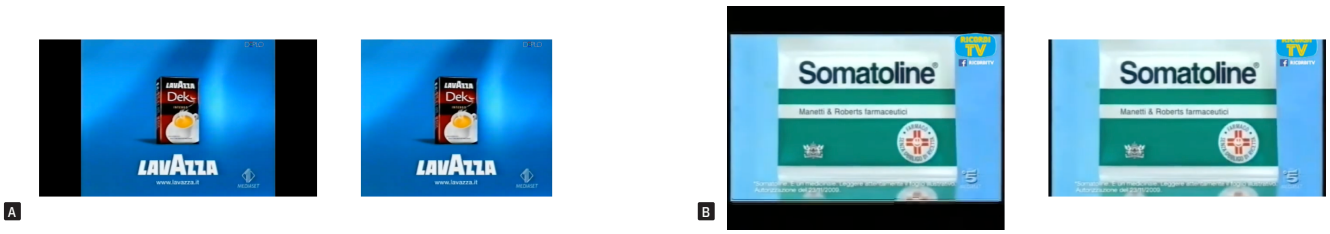


Figure 2. Examples of aspect ratio correction in commercials. (a) Pillar-boxed and (b) letter-boxed formats before and after restoration. Black borders were cropped to recover the original framing and proportions of the footage.

Third, we segmented multi-ad compilation videos (which contained several commercials back-to-back in one file) into individual commercial clips. This segmentation was done programmatically using the time codes annotated by the content creator or manually inserted where missing. Non-commercial fragments (such as television program snippets or transition screens inadvertently included in some uploads) were also removed at this stage. After preprocessing, the final set comprised 10,350 distinct commercials.

We also kept track of the year of broadcast for each commercial. In the majority of cases, this information was explicitly mentioned in the video description or title: we parsed these using regular expressions. For the remaining cases, we performed a manual web search to retrieve missing or incomplete information. Each commercial was also annotated according to the Nice Classification system, an international standard for categorising goods and services developed by the World Intellectual Property Organization (WIPO), a United Nations agency. This allowed us to assign every commercial to a high-level category (e.g., Food & Beverage, Automotive, Toys, etc.). Within these, a finer-grained product type label was applied (for instance, within Food & Beverage: soft drinks, coffee, candy, etc.). These categorisations were obtained from the video descriptions, titles, or inferred from the advertised product. The final metadata table ´ ”Commercials“ contains fields including a unique commercial ID, title, brand, product type, Nice class, release year, lustrum, duration (seconds), aspect ratio, and source. This table contains the master list of commercials in the dataset, which can be linked to all other data records via the commercial IDs.

The distribution of commercials across the Nice Classification categories, as detailed in Table 1, provides a cross-sector coverage of all the categories, with 27 out of 46 total classes containing at least 60 commercials each. This categorical distribution is a direct output of the extraction process from continuous broadcast recordings provided by specialized collectors. By isolating individual commercials from these extensive recordings –specifically by removing non-advertising segments and repeated instances– the resulting frequency of each category serves as a reliable proxy for the actual prevalence and advertising investment within the Italian broadcast market. Consequently, the varying counts across sectors (e.g., the high density in Food & Beverage compared to industrial sectors) are not indicative of collection gaps, but rather faithfully reflect the historical market realities and advertising strategies of the Italian ecosystem from 1980 to 2024.

Scene Segmentation (Shot Boundary Detection)

To analyse the visual narrative structure of each commercial, we segmented videos into scenes (continuous shots) using `PySceneDetect`[8], a Python library designed to detect shot boundaries based on changes in the video content. PySceneDetect identifies abrupt cuts as well as gradual transitions (fades) by analysing frame-by-frame differences, and automatically splits the video into individual scenes. We configured the detector with an adaptive threshold method to handle varied video quality and editing styles. Through experimentation on a subset of commercials, we optimised parameters to avoid over-segmentation due to noisy VHS artifacts or subtle lighting changes. In particular, we set a minimum scene duration of 15 frames (about half a second) to prevent detecting flashing effects as separate scenes. We also instructed the algorithm to skip the very beginning and end of each clip (ignoring the first and last 12 frames) to avoid spurious detections on any initial/final black frames. With these settings, the scene detection reliably splits commercials at true shot changes while ignoring trivial cuts (e.g., a single-frame glitch). Each commercial was thus divided into consecutive scenes, each bounded by the frame indices corresponding to its start and end. On average, commercials in this dataset contain around 5–15 scenes, though the number varies: some simple ads are one unbroken shot, whereas fast-paced ads can have 20 or more quick cuts.

All scene information is stored in a CSV file linked to the commercial ID. As part of scene processing, we extracted a representative video frame for each scene. Specifically, we chose the median frame of the scene to serve as a thumbnail image. This frame is temporarily saved as a PNG image with the same resolution as the video and it is used for extracting colour information. Then it is saved as a WEBP image with a smaller resolution, 180 pixels in height, to store also a

visual reference for the content of that scene. These thumbnails are included in the dataset (organised in folders by commercial ID), enabling quick browsing of scene content without playing the videos.

Figure 3 and 4 illustrates an example of scene segmentation and key frame extraction. By examining scene statistics, we observed that some commercials contained blank scenes at the very start or end consisting mostly of black frames (due to how they were edited or digitised). We implemented a post-processing step to detect and remove such scenes.

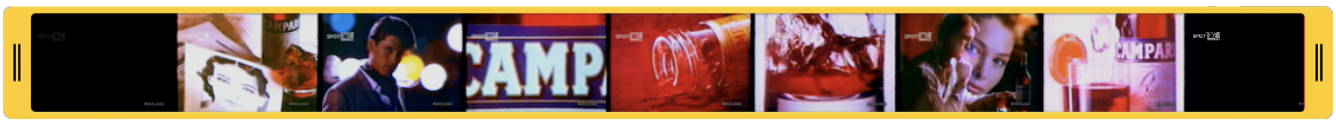


Figure 3. Example of black-frame artefacts in a commercial (Campari, 1990). The sequence shows initial and final frames dominated by black pixels, which were automatically detected and removed from the analysis.

Figure 4. Scene segmentation of the Campari (1990) commercial. The ad was automatically divided into 13 scenes, shown here through the corresponding thumbnail images extracted at the median frame of each segment.

Visual Feature Extraction (Colour Analysis)

We extracted the visual features of each scene through colour analysis. For every representative frame of each scene we extracted a palette of dominant colours that characterise the scene's appearance. We used a clustering-based approach (k-means in colour space), implemented via the `Pylette` library[9], to identify 32 clusters of colours in each representative frame. From each cluster, we extract the medoid colour as representative of that group.

Through this process, the pixels of each frame are grouped into 32 clusters, and the medoid of each cluster serves as a representative colour. Each representative colour is associated with its relative frequency (i.e., the proportion of pixels per cluster). The choice of 32 clusters (k=32) was based on empirical tests: using fewer clusters (around five) did not adequately capture the chromatic diversity of the scenes. In contrast, using much higher values of k produced palettes that were overly detailed and difficult to interpret. Empirically, 32 clusters offered a suitable compromise, providing enough granularity to reflect the visual composition of each scene, while remaining manageable for subsequent analyses. The raw RGB values may be difficult for a human to be interpreted.

Then we further mapped each colour to a perceptually meaningful category by using the `OKLCH`[10] colour space. This is a perceptual colour model, meaning that equal numerical differences correspond more closely to equal visual differences as perceived by the human eye. Unlike CIE Lab, OKLCH was specifically designed to provide smoother lightness and hue perception across the spectrum, making it particularly suitable for both visualisation and computational analysis. Moreover, its adoption in recent standards (e.g., CSS Color Level 4/5) confirms its reliability and growing relevance in practical applications.

ARTICLE IN PRESS

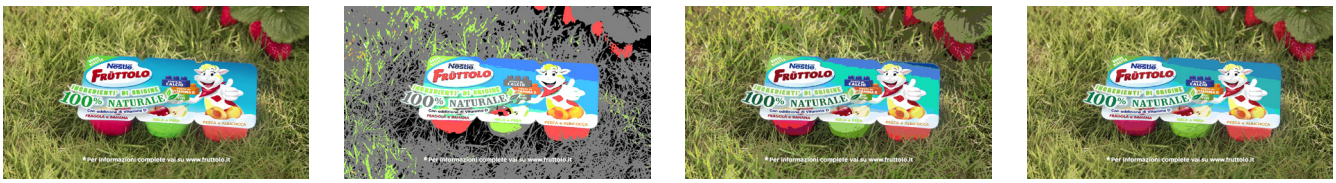


Figure 5. Example of colour quantisation applied to a commercial frame. The original image (a) is compared with its posterised versions obtained by mapping colours to the three hierarchical palettes: (b) `basic` palette with 19 categories, (c) `essential` palette with 311 categories, and (d) `extended` palette with 801 categories. This illustrates the increasing granularity of colour representation across the three levels.

Based on the OKLCH colour space, we propose a hierarchical palette of named colour categories defined at three levels of granularity, with further details provided in the supplementary material.

- **Basic Palette**: 19 broad colour terms, inspired by Berlin and Kay's colour-term hierarchy[11]. These include white, black, gray, red, green, blue, yellow, etc., covering the fundamental colour categories known across languages.
- **Essential Palette**: 311 categories with intermediate specificity.
- **Extended Palette**: 801 categories, capturing very subtle colour differences (for example, dozens of variations of blues and reds).

We mapped each of the dominant scene colours to the closest palette entry using the Coloraide function `closest`[12], which relies on the Euclidean ΔE (CIE 1976) metric[13] to measure colour similarity. The use of this CIE-standard metric ensures that colour differences are computed according to well-established perceptual criteria, while remaining computationally efficient for large-scale processing. Thus, for every scene colour we provide three categorical labels (Figure 5): `closest_color_bas_pal`, `closest_color_ess_pal`, `closest_color_ext_pal`. These labels allow researchers to analyse colour using tag querying, selecting one of the three set of palettes. Thus, for example, it is possible to query "how often the *red* colour appears in ads?", or using more specific colours in the hierarchy, like *crimson* or *pink*.

Derived Colour Metrics

Starting from the colour labels, we computed several derived metrics to summarise colour information. The `frequency within the commercial` measures, for a given colour in a scene, the product of its within-scene frequency by the normalised scene duration, to get a value of how prominent that colour is in the entire commercial. By summing this across all scenes, we can compute a total proportion for each unique colour in the whole ad.

We also computed a term-frequency (`TF`) metric for colours by normalizing these frequencies so that the contributions of multiple occurrences of the same colour are aggregated. To assess the distinctiveness of colours, we applied the well-known `TF-IDF` weighting scheme — commonly used in information retrieval to evaluate how important a term is within a document relative to a collection. In this context, we treat each commercial as a document" and each colour label as a term". The `TF` component captures how frequently a colour appears within a commercial, while the `IDF` (inverse document frequency) down-weights colours that occur across many commercials, highlighting rarer ones.

The `IDF` for each colour was calculated at each palette level (`basic`, `essential`, `extended`) using the total number of commercials N and the number of commercials in which that colour appears. These `IDF` values are provided in separate files (`basic_palette_colors_with_idf.csv`, etc.) listing each standardised colour and its `IDF` score. Colour frequency distributions were normalised by scene length:

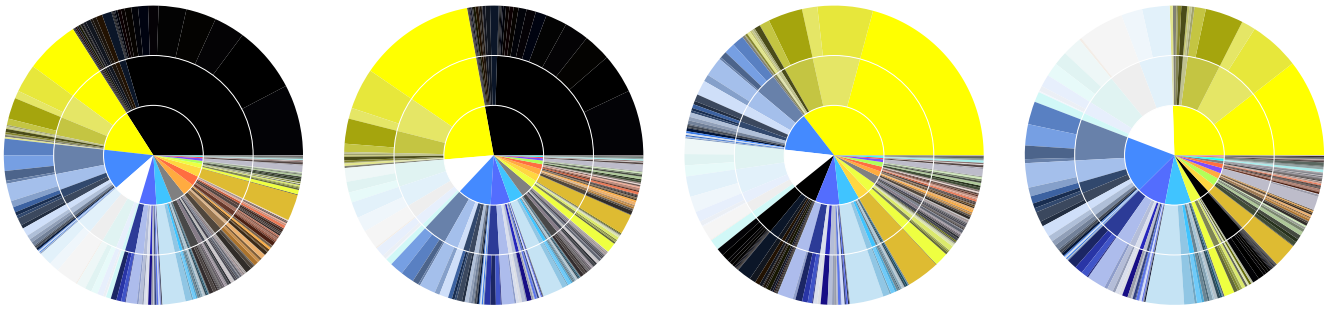
```
frequency_within_the_commercial = frequency_within_the_scene × scene_size_norm
```

The inclusion of `TF` and `IDF` metrics allows users of the dataset to distinguish between ubiquitous colours (like black or white backgrounds, which have high frequency but low `IDF`) and accent colours that are less frequent overall but potentially carry more significance in specific ads (Figure 6).

Audio Feature Extraction

Beyond visuals, the dataset captures detailed audio features for each commercial, facilitating analysis of music, sound effects, and other acoustic aspects. We used the `Librosa` audio analysis library[14] to extract a comprehensive set of features from the audio track of each commercial. These features span several categories commonly used in music information retrieval and audio signal processing:

ARTICLE IN PRESS



A Colours by their colour frequency.

B Colours by their TF-IDF calculated over extended palette.

C Colours by their TF-IDF calculated over essential palette.

D Colours by their TF-IDF calculated over basic palette.

Figure 6. Four colour distributions in the commercial "Poste Italiane". The charts show the hierarchical structure of the colours extracted and then approximated to the extended palette (outer ring), essential palette (middle ring) and basic palette (inner ring).

Feature Categories

Spectral features: We computed spectral centroid, bandwidth, contrast, flatness, and roll-off frequency for the audio. These descriptors summarise the distribution of audio energy across frequency bands (e.g., centroid indicates brightness of sound). They can help distinguish, for instance, a jingle rich in high-frequency instrumentation from a voice-over with most of its energy in lower frequencies.

Timbral features: We extracted 13 Mel-frequency cepstral coefficients (MFCCs) from the Mel-scaled spectrogram. These features capture the spectral shape of the sound, closely related to its timbre. Such representations are useful for characterizing music and for distinguishing music from speech, as they encode a compact "auditory signature" of the audio.

Chromatic features: Using chroma vectors (chroma-STFT, chroma-CQT, chroma-CENS), we summarise harmonic content in terms of the 12 pitch classes (C, C#, D, …), ignoring octave differences. Such features are well-suited for tracking musical key or chord changes in audio containing tonal music, such as commercials with songs or jingles.

Rhythmic features: We applied Librosa's tempo estimator and an external Tempo-CNN model to estimate the beats per minute (BPM) of music segments. We also computed a tempogram, which represents the strength of different tempi over time, allowing us to detect changes in beat rate throughout the clip. These features help characterise rhythmic structure — for example, many commercials use steady background tracks around 120–130 BPM for an upbeat feel, while dramatic ads often opt for slower pacing.

Energy features: We calculated the short-time RMS energy over the commercial, which approximates the loudness envelope (Root Mean Square is a measurement of the average amplitude over a given amount of time). This highlights moments such as loud sound effects or music crescendos.

Other features: We included zero-crossing rate (the rate at which the waveform changes sign, often related to noisiness or percussiveness) and tonal centroid features (tonnetz), which represent harmonic relationships and can capture chordal progression in the audio.

All these features were computed frame-wise (using a hop length of a few milliseconds) and then aggregated or summarised for each commercial. For example, the mean and variance of each MFCC (Mel-frequency Cepstral Coefficients) across the commercials have been stored, or the predominant tempo value has been recorded. The resulting Audio Features table contains one row per commercial, with columns for each feature (a mixture of scalar values and small vectors). In total, several dozen audio feature dimensions are included. By providing pre-extracted audio features, we enable analyses such as comparing the acoustic profiles of commercials over time or by genre. For instance, one could examine whether toys commercials tend to have higher BPM rate than other categories.

Audio Event Classification

A preliminary step before speech transcription was the Audio Event Classification, where, through the YAMNet model[15] from TensorFlow Hub[16], 521 audio event classes were extracted for each commercial. Among these, only the Speech class (e.g., dialogue or voice-over narration) was retained, as all other classes proved statistically irrelevant. Subsequently, only commercials with a Speech confidence score greater than 0.5 were further processed, yielding a final set of 9,793 commercials.

ARTICLE IN PRESS

Speech Transcription and Text Analysis

Many commercials include spoken content – either dialogue, voice-over narration, or slogans – which can be crucial for understanding the message and themes. We generated text transcripts for each commercial's audio speech using OpenAI's Whisper model. In particular, we used the large-v3 model within the `faster-whisper` repository[17], which is trained on multilingual speech recognition.

The model was run on each commercial's audio to produce the transcription in Italian, segmenting the speech into time-coded snippets (sentences or phrases). The output of each commercial is a sequence of transcribed speech segments with their start and end times. We compiled these into a Transcription table where each entry contains the commercial ID and a list of the spoken text segments (along with their timecodes). For example, a 30-second ad might have two spoken lines, each captured as a separate segment in the transcript list.

Transcription segments with average log probability ≤ -1 were discarded. Non-Latin writing systems (e.g., Cyrillic, Devanagari) were detected and filtered out to preserve Italian-language integrity. Manually detected hallucinations and incorrectly transcribed segments were flagged and removed (e.g., 4 non-Italian utterances and 12 implausible segments).

Linguistic Processing

After the transcription process, we proceeded to perform basic text analysis and linguistic annotation. First, we applied a process of lemmatisation and POS (part-of-speech) tagging[18], assigning grammatical categories according to the POS and Universal POS schemes. The Python library used is `Stanza`[19] with the Italian dictionary.

Punctuation marks were excluded from the lemmatisation process, but stop words were not: the reason lies in the fact that in very short texts, such as advertising ones, even a word considered a stopword, e.g., "tutto" ("all") or "più" ("plus", "more") could be significant.

Then, for each lemma (stopwords excluded) within the entire collection of documents, we calculated the term frequency-inverse document frequency (`TF-IDF`), a measure of the term's relative importance in the corpus (N.B. The IDF calculation must be done every time a new set of commercials is added to the collection).

This dataset enables the analysis of term distribution and weight across various commercials, offering a robust foundation for further linguistic or semantic studies.

The lemmas were saved in lowercase, except for those with `UPOS` label set as `PROPN` (proper nouns) and then sorted in alphabetical order.

To preserve the intellectual property of the texts, we provide only the lemmatised form of each term (discarding the original word) and reorder them alphabetically within the dataset.

Linguistic and Psychological Analysis

To complement the morpho-syntactic analysis with standardized psychometric markers, we employed the `LIWC-22` framework (Linguistic Inquiry and Word Count)[20]. `LIWC` is a transparent, dictionary-based tool that quantifies the percentage of words belonging to approximately one hundred pre-validated categories. These range from structural linguistic features (e.g., pronouns, articles) to surface-level proxies for cognitive and emotional processes. For instance, the token "cried" is simultaneously mapped to grammatical categories (past tense verb) and affective markers (negative emotion, sadness), providing a multidimensional signal of the text's emotional tone without relying on opaque inference models.

Using the `LIWC` software (licensed to the University of Pisa), we analysed each transcript to capture dimensions such as emotional tone, cognitive processes, social themes, and prosocial language (e.g., words related to support, sharing, or encouragement). All outputs are linked to the commercials via IDs, enabling integration with visual and audio features. This adds a crucial semantic layer to the dataset, allowing cross-modal studies that connect linguistic messaging with audiovisual elements.

Data Records

The complete dataset is publicly available in the Zenodo repository "Italian TV Commercials Dataset (1980–2024)"[6]. All data records are provided in open formats (CSV for structured data tables, NPZ for Numpy arrays, WEBP for images) to facilitate easy access. The repository is organised into folders by subject (General, Colours, Audio, Text, Scene Thumbnails). A simplified overview of the entities and relationships among the `General`, `Colours`, `Audio`, and `Text` sections is presented in Figure 7.

General

The relational structure of the `General` section is shown in Figure 8.

ARTICLE IN PRESS

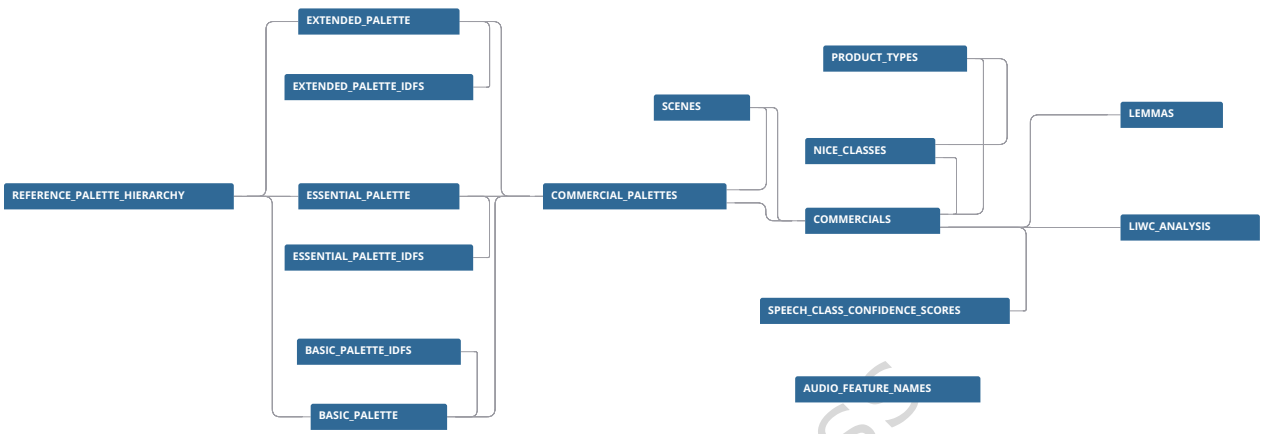


Figure 7. Simplified entity–relationship (ER) diagram of the dataset. The schema summarises the overall structure linking the main data sections — `General`, `Colours`, `Audio`, and `Text` — and their relationships through the shared `commercial_id` key. This high-level overview provides a conceptual map of how the multimodal descriptors are interconnected across the dataset.

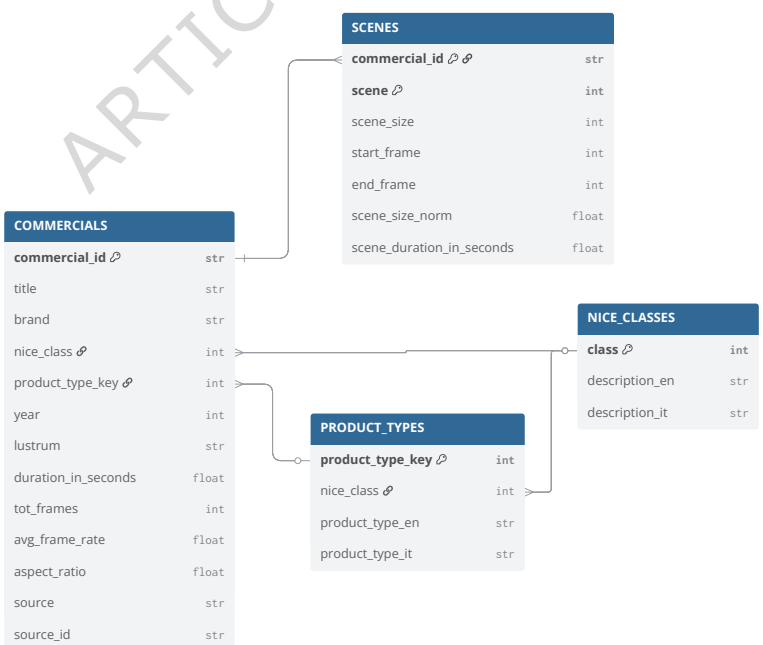


Figure 8. Entity–relationship diagram of the `General` data section. It shows the relational structure connecting commercials with their associated scenes, product types, and Nice classes.

ARTICLE IN PRESS

commercials.csv

The `commercials.csv` data set contains one row per commercial (10,350 rows in total). Each row provides the fundamental information about the advertisement, consisting of the following fields:

- `commercial_id`: the unique identifier of the commercial.
- `title`: the account-assigned video title.
- `brand`: the brand advertised in the commercial.
- `nice_class`: class number, based on industry-standard categories, linked to the `nice_classes.csv` file).
- `product_type_key`: foreign key linking to the `product_types.csv` file.
- `year`: the year of the first airing of the commercial.
- `lustrum`: a five-year period label with pattern `YYYY_YYYY`, where the start year ends with 0 or 5, and the end year ends with 4 or 9 (e.g., 1980_1984, 1985_1989, 1990_1994).
- `duration_in_seconds`: the duration of the commercial in seconds.
- `tot_frames`: the quantity of frames of the commercial.
- `avg_frame_rate`: the average frame rate (fps) of the commercial.
- `aspect_ratio`: the aspect ratio of the video.
- `source`: the source from which the video was retrieved (Spot80 or YouTube).
- `source_id`: the original identifier used by the source (Spot80 or YouTube) to index and retrieve the video.

scenes.csv

The `scenes.csv` data set contains the info about the duration of the scenes detected in each commercial.

- `commercial_id`
- `scene`: the sequential identifier for each detected scene.
- `scene_size`: the scene duration measured in frames.
- `start_frame`: the initial frame number of the scene.
- `end_frame`: the final frame number of the scene.
- `scene_size_norm`: the scene duration normalised.
- `scene_duration_in_seconds`: the duration of the scene in seconds.

nice_classes.csv

The `nice_classes.csv` data set contains 46 classes composed of the 45 classes of Nice classification system plus one extra class (“Associations, organizations, institutions” with the id `-1`).

- `class`: the unique identifier of the Nice class.
- `description_en`: the description in English.
- `description_it`: the description in Italian.

ARTICLE IN PRESS

product_types.csv

The `product_types.csv` dataset provides a further refinement of the Nice classifications by incorporating terms from the Harmonised Database (HDB - https://euipo.europa.eu/ec2/static/html/about-en.html), a common reference of Goods and Services accepted by all EU trademark offices. The HDB contains over 78,000 pre-validated terms available in all official EU languages. It includes the full Nice alphabetical list and additional terms that better reflect current market reality.

Within our dataset, each commercial was manually annotated with one of these refined categories, resulting in 150 unique product types. In practice, the product type can be considered as a subclass of the corresponding Nice class, allowing for more precise identification of the advertised product category.

- `product_type_key`: the unique identifier of the product type.
- `nice_class`: the parent Nice class.
- `product_type_en`: the description in English.
- `product_type_it`: the description in Italian.

Colours

The relationships between the reference palettes and the scene-level palettes are illustrated in Figure 9.

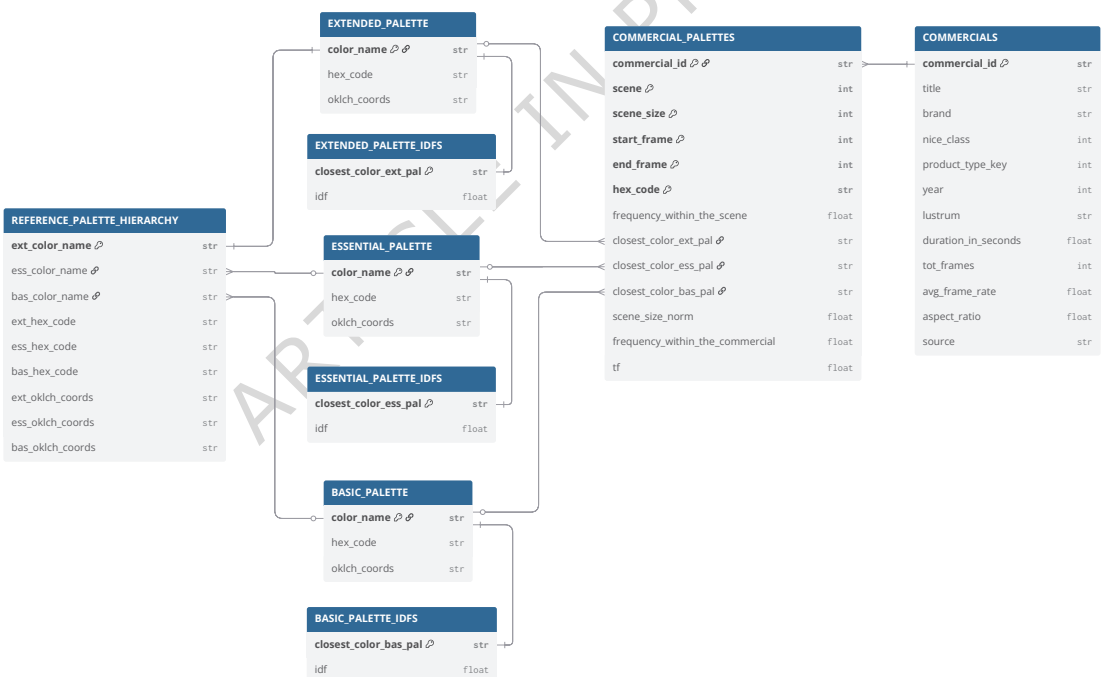


Figure 9. Entity–relationship diagram of the `Colours` section. It details how each commercial is linked to scene-level colour palettes and to the hierarchical reference palettes (basic, essential, extended).

basic_palette.csv

The basic palette made up of 19 colours.

- `color_name`: the name (unique identifier) of the basic colour.
- `hex_code`: the RGB hexadecimal code of the colour in the format `#RRGGBB`.
- `oklch_coords`: the OKLCH coordinates of the colour in the format `(l,c,h)`, i.e., lightness, chroma, hue.

ARTICLE IN PRESS

basic_palette_idfs.csv
The IDF values associated to the basic palette colours.

- `closest_color_bas_pal`: the name (unique identifier) of the basic colour.
- `idf`: the inverse document frequency (IDF) value of the colour.

essential_palette.csv
The `essential` palette made up of 311 colours.

- `color_name`: the name (unique identifier) of the `essential` colour.
- `hex_code`: the RGB hexadecimal code of the colour in the format `#RRGGBB`.
- `oklch_coords`: the OKLCH coordinates of the colour in the format `(l,c,h)`, i.e., lightness, chroma, hue.

essential_palette_idfs.csv
The IDF values associated to the `essential` palette colours.

- `closest_color_ess_pal`: the name (unique identifier) of the `essential` colour.
- `idf`: the inverse document frequency (IDF) value of the colour.

extended_palette.csv
The `extended` palette made up of 801 colours.

- `color_name`: the name (unique identifier) of the `extended` colour.
- `hex_code`: the RGB hexadecimal code of the colour in the format `#RRGGBB`.
- `oklch_coords`: the OKLCH coordinates of the colour in the format `(l,c,h)`, i.e., lightness, chroma, hue.

extended_palette_idfs.csv
The IDF values associated to the `extended` palette colours.

- `closest_color_ext_pal`: the name (unique identifier) of the basic colour.
- `idf`: the inverse document frequency (IDF) value of the colour.

reference_palette_hierarchy.csv
A useful data set for defining the hierarchy of the three reference colour palettes.

- `ext_color_name`: the name (unique identifier) of the `extended` palette colour.
- `ess_color_name`: the name of the `essential` palette colour, parent of the `extended` palette colour.
- `bas_color_name`: the name of the basic palette colour, parent of the `essential` palette colour.
- `ext_hex_code`: the RGB hexadecimal code of the `extended` palette colour.
- `ess_hex_code`: the RGB hexadecimal code of the `essential` palette colour.
- `bas_hex_code`: the RGB hexadecimal code of the basic palette colour.
- `ext_oklch_coords`: the OKLCH coordinates of the `extended` palette colour.
- `ess_oklch_coords`: the OKLCH coordinates of the `essential` palette colour.
- `bas_oklch_coords`: the OKLCH coordinates of the basic palette colour.

ARTICLE IN PRESS

commercial_palettes.csv
The `commercials_palettes.csv` data set contains all the colours extracted (one colour per row) from each scene of each commercial. For each identified colour cluster, we recorded its hex code (six-digit RGB hex value) and the proportion of the frame's pixels that fall into that cluster (frequency within the scene). The collection of those colours and their frequencies constitutes the colour palette for that scene. This information is saved in a per-commercial colour data file (with extension `.pal.csv`), where each row corresponds to one scene-colour combination. In other words, for each scene we have a maximum of 32 entries (rows) listing the colour and its frequency. Together with the scene index, this forms a complete colour profile of the commercial.

- `commercial_id`
- `scene`: the progressive number ID of the scene.
- `scene_size`: the scene duration measured in frames.
- `start_frame`: the initial frame number of the scene.
- `end_frame`: the final frame number of the scene.
- `hex_code`: the hexadecimal representation of the colour.
- `frequency_within_the_scene`: the frequency of appearance in the scene.
- `closest_color_ext_pal`: the closest match of the colour within the `extended` palette.
- `closest_color_ess_pal`: the closest match of the colour within the `essential` palette.
- `closest_color_bas_pal`: the closest match of the colour within the basic palette.
- `scene_size_norm`: the scene duration normalised.
- `frequency_within_the_commercial`: the frequency of appearance in the whole commercial.
- `tf`: the term frequency according to the `closest_color_ext_pal` value.

Audio
The organization of the audio features and the speech presence scores is shown in Figure 10.

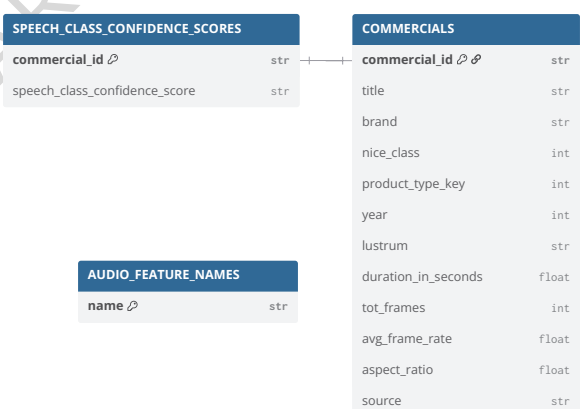


Figure 10. Entity–relationship diagram of the `Audio` section. It illustrates the relationship between commercials and their extracted audio features as well as the speech presence confidence scores.

speech_class_confidence_scores.csv
The data set contains the confidence score of the speech recognition of each commercial.

- `commercial_id`
- `speech_class_confidence_score`: the confidence score for the presence of speech in the commercial.

ARTICLE IN PRESS

audio_feature_names.csv
The data set is just a list of audio feature names, useful to retrieve for each commercial the associated `.npz` files. The pattern of the file path is `audio/features/<feature_name>_files/<commercial_id>.<feature_name>.npz` The 19 audio features are: chroma_cens, chroma_cqt, chroma_stft, librosa_static_tempo_default_prior, librosa_static_tempo_lognorm_prior, librosa_static_tempo_uniform_prior, melspectrogram, mfcc, poly_features, rms, spectral_bandwidth, spectral_centroid, spectral_contrast, spectral_flatness, spectral_rolloff, tempocnn_static_tempo, tempogram, tonnetz, zero_crossing_rate.

- `name`: the name of the audio feature.

Text
The structure of the textual data (lemmas, annotations, and LIWC categories) is depicted in Figure 11.

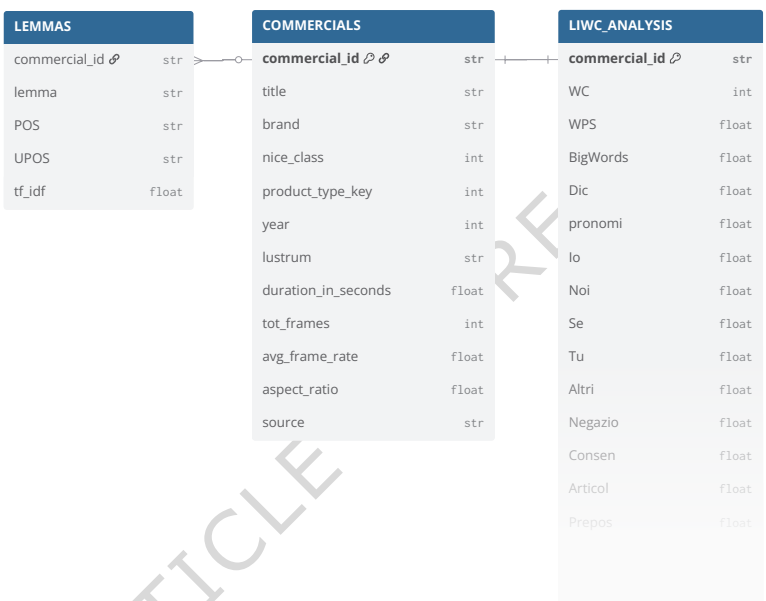


Figure 11. Entity–relationship diagram of the `Text` section. It represents the structure linking each commercial to its lemmas, LIWC features, and linguistic annotations.

lemmas.csv
The `lemmas.csv` provides for each commercial the list of lemmas (alphabetically ordered) of its speech and their part-of-speech tagging. Punctuation marks were excluded from the lemmatisation process, but stop words were not: the reason lies in the fact that in very short texts, such as advertising ones, even a word considered a stopword, e.g., “tutto” (“all”) or “più” (“plus”, “more”) could be significant. Then, for each lemma within the entire collection of documents, we calculated the term frequency-inverse document frequency (`TF-IDF`), a measure of the term’s relative importance in the corpus.

- `commercial_id`
- `lemma`: the lemmatised form of the term.
- `UPOS`: Universal Part-Of-Speech tag, the assigned grammatical category.
- `POS`: language-specific part-of-speech, which is more fine-grained and language dependent.
- `tf_idf`: the relevance value of the lemma in the corpus based on the `TF-IDF` metric.

liwc_analysis.csv
The data set provides for each commercial the values for each of 97 `LIWC` categories. These are organized into a structured metadata table containing normalized scores for both morpho-syntactic elements (e.g., pronouns, prepositions) and psychometric markers (e.g., affective processes such as *anger* or *sadness*). Rather than modeling cognitive processes

ARTICLE IN PRESS

directly, these features serve as surface-level linguistic proxies that enable the quantitative analysis of communicative styles. For instance, a high score in the ‘family‘ category provides a measurable signal of domestic-oriented rhetorical appeals, while variations in ‘tone‘ or ‘clout‘ metrics can help reconstruct the evolution of persuasive strategies over time.

- `commercial_id`
- A series of 97 features (`WC`, `WPS`, `BigWords`, `Dic`, `pronomi`, `Io`, `Noi`, `Se`, `Tu`, `Altri`, `Negazio`, `Consen`, `Articol`, `Prepos`, etc.). The whole list of features is available in the code repository.

Scene Thumbnails (Images)
For each scene listed in `scenes.csv`, an image file is provided showing the scene’s representative frame. These .webp image files are organised in the `thumbnails` directory by commercial. Each commercial has a folder containing its scene images, whose path follows this pattern: `thumbnails/<commercial_id>/<commercial_id>.ss<number_scene>.webp`. The thumbnails have a uniform height of 180 pixels, preserving the original aspect ratio.

Data Overview

To provide a concise overview of the dataset, we report two descriptive summaries. Figure 12 shows the distribution of commercials across time, highlighting the coverage of more than four decades (1980–2024). This temporal span ensures the inclusion of both historic and contemporary advertising materials, enabling longitudinal analyses of changes in style and format. It should be noted that the temporal distribution is not homogeneous: earlier years are comparatively better represented, since the main sources were curated by enthusiasts with a stronger focus on vintage commercials.

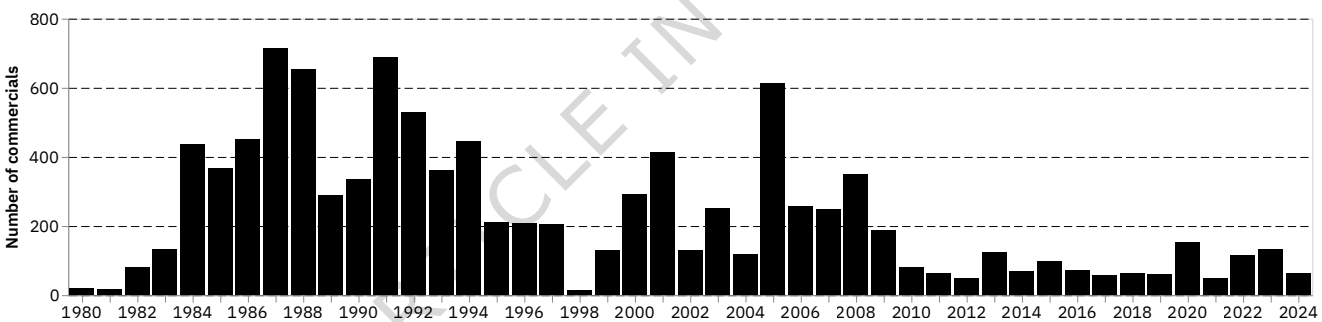


Figure 12. Temporal distribution of commercials in the dataset (1980–2024). The dataset spans over four decades of Italian advertising, with denser coverage in earlier years due to the availability of vintage collections curated by enthusiasts.

Table 1 presents the distribution of commercials by product category, based on the Nice Classification system. The dataset covers a broad range of sectors, with particularly strong representation in Food & Beverage, Automotive, and Toys, reflecting both dominant advertising domains and the diversity of consumer markets represented.

Table 1. Distribution of commercials across Nice classes.

Count	Class
1631	Bleaching preparations and other substances for laundry use; cleaning, polishing, scouring and abrasive preparations...
1600	Coffee, tea, cocoa and artificial coffee; rice; tapioca and sago; flour and preparations made from cereals; bread...
924	Vehicles; apparatus for locomotion by land, air or water
815	Meat, fish, poultry and game; meat extracts; preserved, frozen, dried and cooked fruits and vegetables; jellies, jams...
645	Pharmaceuticals, medical and veterinary preparations; sanitary preparations for medical purposes; dietetic food...
582	Paper and cardboard; printed matter; bookbinding material; photographs; stationery and office requisites...
436	Telecommunications
411	Scientific, nautical, surveying, photographic, cinematographic, optical, weighing, measuring, signalling, checking ...
406	Beers; mineral and aerated waters and other non-alcoholic beverages; fruit beverages and fruit juices...
386	Games, toys and playthings; video game apparatus; gymnastic and sporting articles; decorations for Christmas trees
356	Alcoholic beverages (except beers)
340	Advertising; business management; business administration; office functions; marketing, promotional ...

ARTICLE IN PRESS

Count	Class
316	Clothing, footwear, headgear
189	Insurance; financial affairs; monetary affairs; real estate affairs; investment services...
131	Apparatus for lighting, heating, steam generating, cooking, refrigerating, drying, ventilating, water supply...
127	Precious metals and their alloys; jewellery, precious and semi-precious stones...
124	Services for providing food and drink; temporary accommodation, catering and hospitality services...
114	Machines and machine tools; motors and engines (except for land vehicles)...
94	Hand tools and implements (hand-operated); cutlery; side arms; razors; sharpening instruments...
92	Household or kitchen utensils and containers; cookware and tableware; combs and sponges...
74	Education; providing of training; entertainment; sporting and cultural activities; educational consultancy...
68	Raw and unprocessed agricultural, aquacultural, horticultural and forestry products...
68	Furniture, mirrors, picture frames; containers, not of metal, for storage or transport; home furnishings...
66	Leather and imitations of leather; animal skins and hides; luggage and carrying bags; umbrellas and parasols...
66	Industrial oils and greases; lubricants; dust absorbing, wetting and binding compositions; fuels (including motor spirit)...
62	Surgical, medical, dental and veterinary apparatus and instruments; artificial limbs, eyes and teeth...
60	Transport; packaging and storage of goods; travel arrangement and logistics services...
22	Textiles and substitutes for textiles; bed covers; table covers; household linens...
22	Legal services; security services for the protection of property and individuals; personal and social services...
21	Chemicals used in industry, science and photography, as well as in agriculture, horticulture and forestry; manures...
15	Medical services; veterinary services; hygienic and beauty care for human beings or animals; hospital services...
15	Scientific and technological services and research and design relating thereto; industrial analysis...
14	Associations, organizations, institutions
13	Common metals and their alloys, ores; metal building materials; transportable buildings of metal; railway material...
9	Building construction; repair; installation services
8	Unprocessed and semi-processed rubber, gutta-percha, gum, asbestos, mica and substitutes for all these materials...
6	Ropes and string; nets; tents, awnings, tarpaulins, sails; sacks for the transport and storage of materials...
5	Yarns and threads, for textile use
5	Musical instruments
4	Building materials (non-metallic); non-metallic rigid pipes for building; asphalt, pitch, tar and bitumen; transportable...
3	Carpets, rugs, mats and matting, linoleum and other materials for covering existing floors; wall hangings...
2	Treatment of materials
1	Tobacco; smokers' articles; matches
1	Firearms; ammunition and projectiles; explosives; fireworks
1	Paints, varnishes, lacquers; preservatives against rust and against deterioration of wood; colorants; mordants...

ARTICLE IN PRESS

Technical Validation

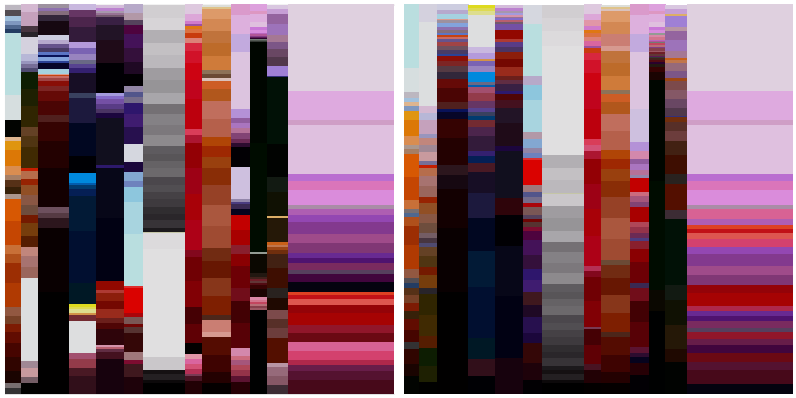
We undertook several measures to ensure the technical quality and consistency of this dataset. Data integrity and completeness were verified through automated checks. At the end of the processing of each commercial video, we checked the integrity and completeness of the values in the data collection.

Video segmentation validation

The scene detection procedure was tuned to improve boundary precision and avoid artifacts such as black frame overlays. We excluded the first and last 12 frames of each commercial to prevent the inclusion of non-content frames (e.g., transitions or padding from compilation videos). Additionally, a black frame filtering strategy was applied: first and last scenes with more than 90% black pixels –identified via perceptual colour mapping to the `black` label in the `essential` palette – was excluded from the colour analysis to avoid encoding technical artifacts instead of actual commercial content.

Scene Detection Validation

We manually reviewed a random sample of 100 commercials, comparing the original video to the extracted data. This review confirmed that scene boundaries were correctly identified (with a precision of ~98% for cut detection as judged by human inspection) and that no major scene was missed. We worked on the fine-tuning for the PySceneDetect library, by evaluating both ThresholdDetector() and AdaptiveDetector() strategies, eventually choosing the latter with default parameters) and repeated the operation until most of the transitions in the test videos were captured with minimal false cuts.



A Campari Colours sorted by hue.

B Campari Colours sorted by lightness.

Figure 13. Marimekko charts of colour composition for the Campari (1990) commercial. Each column represents a scene, with width proportional to its duration and vertical segments showing the relative frequency of the 32 colours detected. Colours are ordered by (a) hue and (b) lightness, illustrating chromatic variation across the ad's temporal structure

Colour Analysis Validation

To validate the robustness of the extracted colour palettes, we implemented a form of visual analytics, comparing them with the corresponding frame images. For example, in a *Poste Italiane* commercial visually dominated by blue and yellow, the palette file for the relevant scene correctly identified these two hues as the most frequent 6, in line with perceptual expectations. Furthermore, the application of the `TF-IDF` metric reduces the relative weight of background colours, such as black and dark greys, thereby highlighting the accent colours that shape the visual identity of the commercial. To this end, we conducted a series of visual inspections using custom Marimekko-style charts (Figure 14). In these charts, each commercial is represented as a sequence of adjacent vertical bars, one for each scene. The width of each bar is proportional to the scene's duration in frames, while the height of each coloured segment within a bar reflects the relative frequency of that colour in the corresponding scene's palette. This visualisation technique allows for an intuitive inspection of both the temporal structure and the chromatic composition of each commercial. Longer scenes appear wider, while more frequent colours occupy larger vertical spans within their scene block. For instance, scenes featuring natural landscapes (e.g., skies, meadows) reliably show mostly blue and green blocks.

These plots reveal high internal consistency across scenes and confirm that the extracted colours align well with human visual perception. They also make it possible to spot anomalies or outliers (e.g., scenes dominated by total black frames with a yellow logo). Overall, the visual features derived from the colour analysis provide a robust and interpretable quantitative profile, a sort of footprint, of each commercial's visual design, capturing both the pacing of visual cuts and the palette of colours employed throughout its duration.

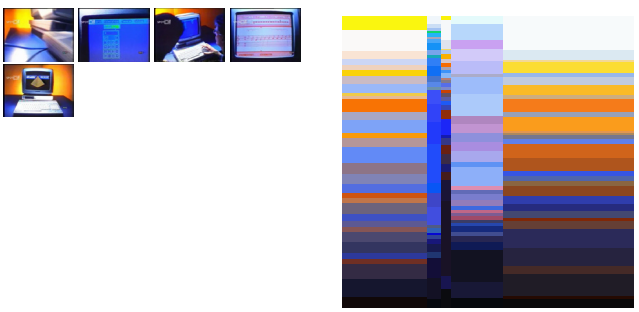
Marimekko charts can be customised to prioritise different dimensions of colours features. For example, Figure 13 shows the colours of the same commercial, using two different strategies of sorting: in the right one, the colours in each column are sorted by hue; in the left one, each column shows the colours sorted by lightness.

We checked the impact of the colour hierarchical palettes, by comparing the resulting term-frequency at each level of the taxonomy. For example, in Figure 6 we have a visual summary of this check. Figure 6a uses the frequencies of the *Extended palette* tags: the dark colours cover almost one third of the whole set of pixels, mainly due to the black bars of the letterboxing of the spot (we intentionally left those bars). Figure 6b shows the `TF-IDF` value for each *Extended palette* tag: here the impact of darker colour is reduced. Figure 6c computes the `TF-IDF` at the level of the *Essential* palette: here we can sense the benefit of having a more compact set of labels that help in reducing the noise of repeated colours. This level of abstraction of colour labels proved to be the most effective. Figure 6d uses the *Basic* palette: here we loose many detauils of the overall palette, but it highglights the distinguishing colour of the brand.

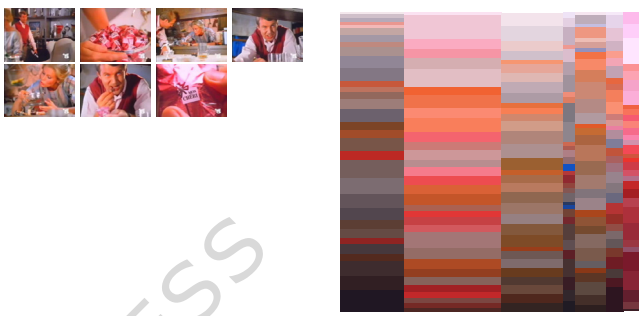
These visual summaries helped in verifying the consistency of the colour data across the commercials, since it is possible to check at a glance how the frequency measures aligns with the visual perception.

ARTICLE IN PRESS

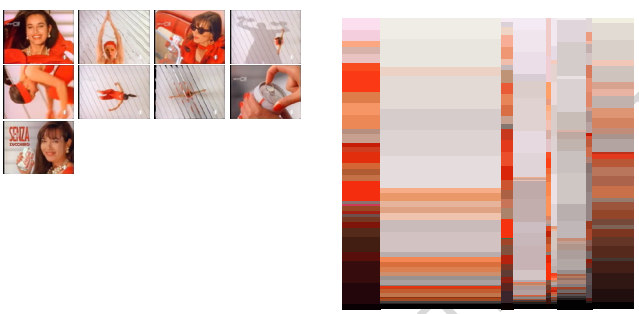
A Computer Olivetti (1986).



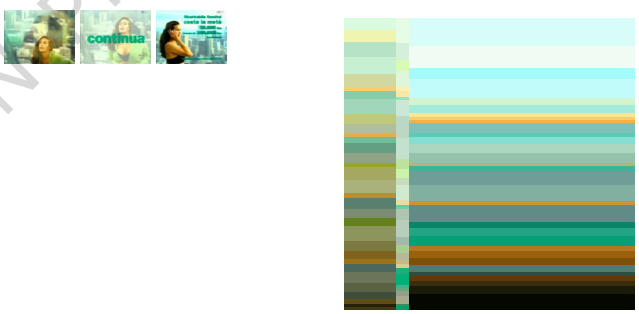
B Mon Chéri Ferrero (1988).



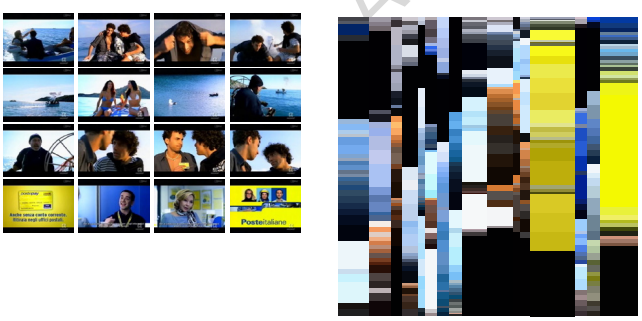
C Coca-Cola light (1991).



D Omnitel Vodafone (2001).



E Poste Italiane (2003).



F Barilla (2013).

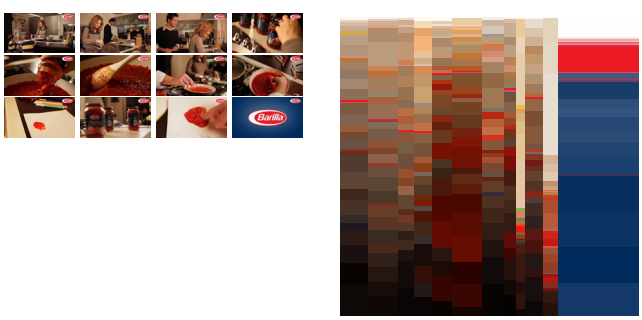


Figure 14. Visual summaries of six Italian TV commercials (1986–2013). Each pair shows scene thumbnails (left) and the corresponding Marimekko colour distributions (right), illustrating stylistic and chromatic patterns across brands and decades: (a) *Computer Olivetti* (1986), (b) *Mon Chéri Ferrero* (1988), (c) *Coca-Cola light* (1991), (d) *Omnitel Vodafone* (2001), (e) *Poste Italiane* (2003), and (f) *Barilla* (2013).

ARTICLE IN PRESS

Audio Feature Validation

When validating audio features, the accuracy of the extracted data is limited by the quality of the source recordings. That quality is largely set by two factors: the strength of the broadcast signal as it reaches the user's antenna, and the recording equipment used over time.

Before extracting audio features, we first normalised the recordings. This involved converting the MP4 files to WAV, downmixing the audio to mono, and resampling to 16kHz. The process guarantees that every file has the same format, channel configuration, and sampling rate, which makes the subsequent feature extraction more reliable and easier to compare.

We also verified that our feature definitions match standard practice by using well-established libraries such as Librosa and published CNN models, along with parameters that have been reported in the literature. This ensures that researchers can trust the extracted descriptors as a solid baseline.

Transcription and Annotation Validation

Speech segments were transcribed using the Whisper large-v3 model via `faster-whisper`.

The resulting transcripts were carefully reviewed for quality. The Whisper model has been shown to achieve near-human level accuracy on many languages, and we found that for clearly spoken ads the transcription was very accurate (including correctly handling Italian names and product terms). In cases of background noise or heavy music, minor errors were possible; we manually spot-checked a sample and corrected a few obvious mis-recognitions, especially where they could affect downstream text analysis (for instance, if a brand name was misheard).

We took care to ensure the text data's consistency by using a single tool for NLP preprocessing across the entire corpus, and by manually verifying that encoding issues (accented Italian characters, etc.) were handled correctly. The Italian language processing is inherently supported by the chosen tools (Whisper), and we confirmed that domain-specific words (like product names) were preserved in transcripts, adding them as hot-words, even if they are out-of-vocabulary for the language model.

To comply with copyright and fair use constraints, full transcripts of the commercials are not included in the dataset. Instead, the textual layer is released in the form of tables containing the part-of-speech (POS) annotations (made with Stanza) for each word, along with lemmatised forms and grammatical features. This representation retains the structural and stylistic properties of speech without exposing verbatim content.

To further ensure that original scripts could not be reconstructed from the POS data, we conducted an adversarial test using large language models (LLMs). A significant sample of transcripts was selected, focusing in particular on cases that appeared most vulnerable to potential reconstruction (e.g., very short texts). The models were then prompted with POS-tagged sequences and asked to regenerate the original spoken text. In all checked cases, the outputs were generic and uninformative.

These combined strategies ensure that the linguistic data is ethically compliant, non-reconstructive, and still analytically useful.

Dataset Consistency

Finally, we ensured internal consistency across the entire dataset. Each commercial has a unique ID used in all tables; we wrote a script to verify that there are no missing or duplicate IDs in any table.

The dataset's completeness was also evaluated: the collection of 10,350 commercials is large, yet we acknowledge it may not include every ad aired in Italy in the period. However, the focus was on compiling a representative and diverse sample. The distribution by year and category suggests we have broad coverage, with at least dozens of examples in each major category and lustrum.

In summary, the technical validation gives us confidence that the dataset is reliable for reuse. The combination of manual inspection and automated checks has minimised errors and ensured that the data truly reflects the content of the commercials.

Usage Notes

This dataset[6] is intended to be as self-contained and accessible as possible. All files are in human-readable formats, and the repository includes documentation and example code. Researchers can load the CSV files using standard data analysis software (Python pandas, R, Excel, etc.) and link them via the `commercial_id` field.

A detailed documentation and demo code to replicate the proposed processing pipeline are available in the GitHub repository.[21]

ARTICLE IN PRESS

Key Considerations

Data integration: The multi-modal nature means that users can combine features for richer analysis. For example, one could examine correlations between visual and textual features (do commercials that use a lot of the colour pink also use more words related to family?). To do this, one would join the relevant tables (colour and text) on `commercial_id`.

Temporal analysis: Given the multi-decade span and the multiple sources, users should be cautious of changes in broadcasting and recording technology over time (there are commercials recorded on VCRs and then digitised, and commercials that are originally digital).

Colour data usage: The colour palette extracted for each scene can be visualised directly or used in aggregated form, depending on the research objective. To enable flexibility while maintaining interpretability, each colour cluster derived from the frame is mapped univocally across a three-level perceptual hierarchy:

- The *Extended palette* includes over 800 fine-grained colour descriptors, capturing subtle chromatic nuances. This level is suitable for computer vision tasks such as image clustering or style recognition, where pixel-level precision matters.
- The *Essential palette*, comprising 311 human-readable colour labels, strikes a balance between perceptual granularity and semantic accessibility. This level is particularly effective for statistical modeling and weighting, such as applying `TF-IDF` to distinguish salient colours across commercials without overfitting to near-identical shades.
- The *Basic Palette* consists of 19 universally recognizable colour categories (e.g., red, blue, yellow, black). Its primary advantage lies in semantic interpretability: each colour is explicitly named and easily understood, making it ideal for communication studies or visual storytelling analyses.

Since the mapping from `extended` to `essential` to `basic` is univocal and hierarchical, researchers can seamlessly shift between levels of detail without introducing ambiguity or conflicting classifications. For example, a scene-dominant colour identified as #b39d4f (an amber tone) will map consistently to `amber_l70_c11_h93` in the `extended` palette, to `amber_l60_c7_h93` in the `essential` palette and to `amber` in the basic palette. This hierarchical structure supports both exploratory and comparative work across disciplines and decades of advertising styles. A detailed description of the colour vocabulary used in the dataset can be found in the supplementary material.

Audio data usage: The audio feature vectors provided can feed into models for tasks like genre classification or to find patterns. Users should note that some features (like MFCCs) are given as aggregated summaries; if frame-level analysis is needed, one would have to extract features from the original audio files (not included in the distribution). We also caution that the audio features from very short ads may be less reliable statistically.

Lemma data usage: The lemmas are in Italian. Non-Italian-speaking researchers might need to translate or at least be aware of the language context. Automatic translation can be applied to the lemmas if needed, but for many analyses it may not be necessary. If performing NLP tasks, we recommend using Italian-specific resources since the text is Italian.

Social Sciences: Our dataset results from a modular, bottom-up computational architecture designed to augment, rather than replace, semiotic and sociological analysis. By adopting a data-driven approach, we provide a suite of quantifiable features—ranging from hierarchical color palettes and scene durations to LIWC psychometric markers—that serve as agnostic "material ingredients" for high-level interpretation. While raw pixel values lack inherent semantic meaning, our pipeline transforms these signals into computational proxies for visual analysis. We bridge this gap by abstracting the color space into standardized linguistic labels (mapped to the 19 Berlin and Kay categories), moving from signal processing to human-readable descriptors. To identify rhetorical significance, we apply TF-IDF to scene-level color histograms. This allows researchers to mathematically distinguish generic background tones from distinctive "accent colors." For instance, this metric can isolate specific branding signatures—such as the yellow and blue identity of Poste Italiane (Figure 6)—enabling the automated retrieval of brand patterns across decades without requiring prior manual training.

This resource is intended as an adaptable framework for diverse analytical directions across multiple domains. By releasing the full preprocessing codebase, we ensure transparency and allow researchers to extend, replicate, and adapt the pipeline. Notably, our approach involves no explicit learning phase or model training, minimizing the risk of amplifying inherent data biases. The resulting dataset serves as a flexible infrastructure for future scholarly work. Rather than offering a "turnkey" solution, it provides the raw material for domain experts to layer their own annotations, theoretical frameworks, and interpretive lenses. Researchers can refine our "surface-level" proxies by integrating specific theories of semiotic modes and patterns ([22]), treating the collection as a testbed for longitudinal studies in communication dynamics, cultural evolution, or media history.

ARTICLE IN PRESS

Licensing and Ethics

The dataset is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits reuse with appropriate attribution. All content consists of advertisements originally intended for public broadcast, and the dataset contains no personally identifiable information.

Our research methodology and data publication approach have been carefully designed to comply with European Union and Italian legal frameworks while ensuring ethical data usage. This section outlines the key legal and ethical considerations governing our work.

Research fair use and data mining rights

Our data analysis falls within the robust legal framework for Text and Data Mining (TDM) for scientific research, as articulated in EU Directive 2019/790 Article 3 and Italian Copyright Law Article 70-ter. This framework explicitly permits automated computational techniques to analyze large-scale digital corpora—encompassing text, audio, and visual metadata—to uncover underlying patterns and correlations that would otherwise remain inaccessible.

Our methodology is specifically designed to fulfill these legal mandates by prioritizing a transformative approach to data processing. Rather than reproducing or redistributing original content, our work focuses exclusively on deriving aggregate patterns and metrics. This transition from raw media to structured data is supported by security protocols maintained throughout the pipeline, ensuring that the processed raw data remains secure. By focusing on these derived patterns, the dataset functions as a factual research tool that respects the spirit of TDM regulations while maximizing the utility of the data for the scientific community.

Data processing and storage

To ensure full legal compliance and ethical stewardship, our methodology integrates several **critical safeguards** within a "privacy-by-design" workflow. Central to this approach is the use of **automated analysis techniques**, which extract research-oriented features (such as color palettes, linguistic markers, and audio characteristics) while effectively abstracting the information from its original broadcast context. Crucially, these analysis outputs are specifically engineered to **prevent the reconstruction** of the source content, shifting the focus toward the derivation of **aggregate patterns and quantitative metrics** that support scientific inquiry rather than media consumption. This transformative process concludes with a definitive measure of data security: the **original video materials are systematically deleted** immediately after feature extraction. By retaining only the derived analytical data and low-resolution representative thumbnails, we establish a clear boundary between the raw copyrighted material and the factual, non-infringing research dataset.

Content analysis framework

Our analytical framework is fundamentally **transformative and research-focused**, designed to convert media into structured data. Central to this process is the color analysis, which generates rigorous **quantitative metrics** through a custom-built color system, rather than relying on proprietary models, as detailed in the supplementary material. This commitment to data abstraction extends to the audio and linguistic layers: audio processing extracts technical features at quality levels specifically engineered to **prevent reconstruction**, while text analysis focuses on producing **derived metrics**, such as sentiment scores, rather than full transcriptions. By prioritizing the generation of derived representations, the pipeline transforms raw media into a structured suite of insights, fostering novel hypotheses and diverse analytical perspectives.

Legal basis for publication

Our publication approach is supported by *Article 70 paragraph 1bis of Italian Copyright Law*, which permits free publication of content for scientific purposes. Additionally, our research qualifies under *GDPR Article 89*'s provisions for scientific research, with appropriate safeguards implemented.

No personally identifiable information is present; adverts occasionally feature actors or public figures, but since these are published ads, analysing them for research is generally permissible under transformative use. Nonetheless, researchers should handle the data ethically, especially if performing analyses that could be sensitive.

The research nature of our analysis is further supported by our methodological approach: we analyse advertising content through **systematic analytical procedures**, and make the derived data available **exclusively for research purposes**. This enables scholars from various disciplines, including social scientists, to conduct research while maintaining appropriate legal and ethical safeguards.

Usage restrictions

The data is made available exclusively for **research and academic purposes**, thus **commercial use is not permitted**. This distinction is reinforced by the nature of the data itself: both the methodology descriptions and the resulting

ARTICLE IN PRESS

quantitative measurements represent **factual information** that is inherently not subject to copyright. By establishing these parameters, we ensure that the dataset remains a dedicated resource for scholarly advancement, contributing to the global body of knowledge while rigorously respecting intellectual property rights and data protection requirements.

Computational Requirements

The dataset in total is moderately large. Processing the entire dataset at once will require a machine with sufficient memory and storage. However, typical analyses should be feasible on a standard modern laptop by loading the CSVs. We provide code in the repository to demonstrate reading and manipulating the data efficiently.

In case users have questions or encounter issues, we have provided contact information in the repository. Future updates may include additional data or improvements. We welcome collaborators who want to extend this dataset, for example by adding annotations – such additions can be shared under the same open terms.

Data Availability

The dataset is openly available on Zenodo at https://zenodo.org/records/19100800.

Code Availability

The documentation and code required to reproduce the data collection and feature extraction methodology are available on GitHub at https://github.com/kdd-lab/tv-commercials-dataset.

The repository contains Python scripts corresponding to each major step described in the Methods.

The code repository includes scripts for:

- `1_color_and_thumb_extraction.py` – Scene, colours and thumbnails extraction
- `2_ref_palette_idf_calculation.py` – IDF calculation
- `3_audio_feature_extraction.py` – Audio feature extraction
- `4_audio_transcription_and_lemmatization.py` – Audio transcription and lemmatisation

The exact version of each external library is specified in a `requirements.txt` file in the code repository. The code repository is organised to facilitate reproducibility: one can follow the README instructions to rerun the pipeline on the original raw data. The code includes parameter settings (e.g., the threshold values for scene detection, the clustering parameters for colour extraction, etc.) exactly as used to generate the published data. We encourage users to use these scripts as a starting point for their analyses or to adapt them for similar projects All code is released under an open-source license (Apache2 License), allowing modification and reuse. No proprietary software was used in generating this dataset – all steps rely on open-source tools, ensuring that anyone can replicate or audit the processes without needing special software.

ARTICLE IN PRESS

Funding

This work has been supported by SoBigData.it, which receives funding from the European Union – NextGenerationEU, under the Italian National Recovery and Resilience Plan (PNRR), Project “SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics” (Prot. IR0000013, Avviso n. 3264 del 28/12/2021).

Acknowledgements

The authors are grateful to Michele Logrippo, curator of Spot 80, for generously providing access to the archival commercial footage that constitutes a crucial component of this dataset. We also wish to thank Virginia Morini for her contribution to the extraction of LWIC features from the text, Andrea Fedele for his support on the audio analysis, and Roberta Savella for her guidance on legal aspects. Finally, we would like to express our sincere appreciation to Giovanni Mauro for his insightful feedback and valuable advice throughout the development of this work.

ARTICLE IN PRESS

Authors Contribution

Daniele Fadda conceived the study, designed the overall structure of the manuscript, and co-developed the methodological framework. Giorgio Bellate was responsible for the technical implementation, including data collection, preprocessing, and feature extraction, and contributed to the methodology. Salvatore Rinzivillo provided scientific supervision throughout the project, contributed to the conceptual and methodological development, and critically revised the manuscript for important intellectual content. All authors discussed the results, contributed to the final manuscript, including images and approved its submission.

References

1. Campos, D. et al. The nutritional profile of food advertising for school-aged children via television: A longitudinal approach. Children **7**, 230, 10.3390/children7110230 (2020). Open access.
2. Emond, J. A. et al. Influence of child-targeted fast food tv advertising exposure on fast food intake: a longitudinal study of preschool-age children. Appetite **140**, 134–141, 10.1016/j.appet.2019.05.012 (2019).
3. Tao, L. et al. A large-scale television advertising dataset for detailed impression analysis. Multimed. Tools Appl. **83**, 18779–18802, 10.1007/s11042-023-14704-7 (2024).
4. Sagar, A., Srivastava, R., R T, R., Kesav Venna, V. & Sarvadevabhatla, R. K. MAdVerse: A Hierarchical Dataset of Multi-Lingual Ads from Diverse Sources and Categories. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 8072–8081, 10.1109/WACV57701.2024.00790 (2024). ISSN: 2642-9381.
5. Barari, M. & Eisend, M. Computational content analysis in advertising research. J. Advert. **53**, 681–699, 10.1080/00913367.2024.2407642 (2024). Published online 15 October 2024.
6. Fadda, D., Bellante, G. & Rinzivillo, S. Italian TV Commercials Dataset (1980–2024), 10.5281/zenodo.19100800 (2025).
7. Schraulek, M. L. M. N. P. Spot 80 (2024). Accessed: 2025-02-15.
8. Castellano, B. Pyscenedetect (2025). Video Cut Detection and Analysis Tool. Repository: https://github.com/Breakthrough/PySceneDetect.
9. Stangeby, I. Pylette, 10.5281/zenodo.14757253 (2025).
10. Ottosson, B. A perceptual color space for image processing. https://bottosson.github.io/posts/oklab/ (2020). Accessed: 2025-09-30.
11. Berlin, B. & Kay, P. Basic Color Terms: Their Universality and Evolution (University of California Press, 1969).
12. Muse, I. ColorAide: A Python Library for Modern Color Space Management (2025). Version 5.1 released August 31, 2025, available via PyPI.
13. Muse, I. Color . delta_e. https://facelessuser.github.io/coloraide/api/#delta_e. Accessed: November 2025.
14. McFee, B. et al. librosa: Audio and music signal analysis in python. In Huff, K. & Bergstra, J. (eds.) Proceedings of the 14th Python in Science Conference, 18–25 (2015).
15. Google Research. YAMNet on TensorFlow Hub. https://tfhub.dev/google/yamnet/1 (2019). Accessed: 2025-04-01.
16. Abadi, M. et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems (2016). 1603.04467.
17. faster-whisper contributors, S. . faster-whisper: Faster whisper transcription with ctranslate2. https://github.com/SYSTRAN/faster-whisper. Version as of access, MIT License; accessed: 2025-09-30.
18. Petrov, S., Das, D. & McDonald, R. A universal part-of-speech tagset. In Chair), N. C. C. et al. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (European Language Resources Association (ELRA), Istanbul, Turkey, 2012).
19. Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. Stanza: A python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 101–108, 10.18653/v1/2020.acl-demos.14 (Association for Computational Linguistics, 2020).
20. Boyd, R. L., Ashokkumar, A., Seraj, S. & Pennebaker, J. W. The development and psychometric properties of liwc-22. Austin, TX: Univ. Tex. at Austin **10** (2022).
21. Fadda, D., Bellante, G., Rinzivillo, S. & et al. Italian tv commercials dataset – methodology demo pipeline. GitHub repository (2025). Version at 2025-10-30; urlhttps://github.com/kdd-lab/tv-commercials-dataset.

ARTICLE IN PRESS

22. Hiippala, T. Distant viewing and multimodality theory: Prospects and challenges. Multimodality & Soc. **1**, 134–152, 10.1177/26349795211007094 (2021). https://doi.org/10.1177/26349795211007094.

ARTICLE IN PRESS