Advanced Sentiment Analysis of YouTube Product Reviews using Multimodal AI

A System for Deep Analysis and Synthesis of Video-Based Product Opinions to Inform Market Strategy and Development

Project Report

GABRIELE GIUDICI
GABRIELE BILLI CIANI

University of Pisa
MSc. in Artificial Intelligence and Data Engineering
Business and Project Management Course
Academic Year 2024–2025

SCUOLA
DI INGEGNERIA

Contents

GitHub Repository

The GitHub repository containing all the project files is available at:

https://github.com/gabrielebilliciani/YT-multimodal-sentiment-analysis

In various parts of the current report we refer to the GitHub repository for additional details. The README file explains where to find the complementary material, such as the prompts used for the multimodal analysis, the JSON files containing the results of the analysis, and the configuration files used to run the experiments.

1 Introduction

The proliferation of digital platforms has paved the way for an era where user-generated content (UGC) is not merely a byproduct of online interaction but a critical data source for businesses worldwide [1, 2]. This content, ranging from textual posts to intricate video productions, offers an unfiltered window into consumer preferences, product satisfaction levels, emerging market trends, and competitive landscapes. Companies are increasingly leveraging this data to refine product development cycles, design targeted marketing campaigns, and improve customer relationship management strategies. Among the diverse forms of UGC, video reviews produced and uploaded by influencers on platforms such as YouTube have gained exceptional importance, emerging as a highly influential factor in shaping consumer perceptions and guiding purchasing decisions [3]. These video reviews present a complex, multimodal narrative, rich with information packed with spoken language, visual presentation of products, and the subtle emotional expressions of the reviewer.

The richness of video content, however, introduces significant analytical complexities. Traditional sentiment analysis methodologies, which are predominantly text-based, while well-established and valuable [4, 5], can only capture a partial view of the information spectrum available in a video. A truly holistic understanding requires the capability to interpret a mixture of signals: visual cues, such as facial expressions, body language, and product interaction; audio elements, including the tone of voice, stressing some particular words or sentences, and speech inflections; and the explicit textual content from transcriptions or on-screen text [6, 7, 8]. This imperative for comprehensive analysis drives the need for advanced multimodal approaches.

The recent emergence of powerful Generative Artificial Intelligence (GenAI) models, marks a paradigm shift in AI capabilities. These models are distinguished by their proficiency in processing, understanding, and generating content across a diverse array of data types, including text, images, audio, and video. This multimodal ability opens new and original opportunities for handling the analytical challenges posed by complex video content. Our project is dedicated to the development of an innovative system that capitalises on the capabilities of recently released GenAI systems (in particular we focus on Google's Gemini) to perform an in-depth, multimodal analysis of YouTube influencer product reviews. The primary strategic objectives of this system are:

1. To provide businesses with a comprehensive and all-round overall sentiment and affective summary concerning their products, as articulated and expressed by YouTube influencers and/or generic review videos.

2. To offer a detailed comparative analysis, benchmarking a company's products against those of its direct competitors, based on complex influencer feedback.

3. To achieve this by moving beyond mere textual analysis, integrating insights derived from the visual (e.g. influencer's facial expressions, product demonstration quality) and auditory (e.g. tone of voice, enthusiasm levels) dimensions of the video reviews.

Before coming to the implementation and results of the system, the initial part of the current report is structured as a foundational introduction to our project. It aims to contextualise our work within the established and evolving domains of sentiment analysis, affective computing, and multimodal artificial intelligence. We will undertake a targeted review of the current state-of-the-art, drawing upon a curated selection of recent and relevant scientific literature directly informing our approach. This review will highlight existing capabilities, prevalent methodologies, inherent challenges, and future directions in the analysis of emotions and sentiments, with a particular emphasis on multimodal sources such as YouTube videos, highlighting the novelty and potential impact of our proposed GenAI-driven system.

2 State-of-the-Art in Sentiment and Affective Analysis

The desire to computationally understand and interpret human emotion and opinion from digital content has been a consistent and driving force in computer science research. This task holds deep implications across a multitude of business applications, from market research and brand management to customer service and product innovation. The fields of sentiment analysis and affective computing are a crucial aspect of this pursuit, providing the theoretical frameworks and practical tools for such analysis.

2.1 Sentiment Analysis

Sentiment analysis, or opinion mining, is primarily concerned with the systematic identification, extraction, quantification, and study of affective states and subjective information using natural language processing (NLP), text analysis, computational linguistics, and biometrics [3]. While early efforts focused on lexicon-based approaches, the field has significantly advanced with machine learning and deep learning techniques [1, 4]. The emergence of Large Language Models (LLMs) has further revolutionised text-based sentiment analysis, offering sophisticated understanding of context [4]. However, challenges such as detecting implicit sentiment, handling sarcasm, and mitigating biases persist [5]. The analysis of specific content types, like hate speech in YouTube comments, also remains an active area of research [9].

2.2 Affective Computing

Affective computing extends beyond simple polarity to incorporate a broader spectrum of human emotions and affective states [10]. This field explores emotion recognition, sentiment analysis, and even personality assignment using various AI techniques, including LLMs and multimodal approaches. Applications are diverse, ranging from AI chatbots and mental health support to safety systems [10, 11]. The interpretation of non-verbal cues is a key aspect, as highlighted in studies on older adults [12] and the development of multimodal datasets like AFFEC [11]. Ethical considerations are also paramount when dealing with sensitive emotional data [10].

2.3 Multimodal Emotion and Sentiment Analysis

Recognising that communication is inherently multimodal, research has increasingly focused on integrating information from various channels such as text, audio, and video, in order to achieve a more accurate and comprehensive understanding of affect and sentiment [7, 3, 8].

Analysing video content, such as YouTube reviews, inherently demands such a multimodal approach. Visual signals as facial expressions are powerful indicators of emotion [6], and auditory signals as the tone of voice can significantly alter meaning. The MMLA benchmark by Zhang et al. [13] and the DEEMO framework by Li et al. [14] are examples of efforts to evaluate and enable multimodal understanding, including in privacy-preserving contexts. Liu et al. [15] specifically investigated Multimodal LLMs' (MLLMs) ability to grasp visual concepts in YouTube Shorts. The effective fusion of these heterogeneous data sources remains a key challenge [1, 7, 3].

3 The Proposed System: A Generative AI Approach

The methodologies discussed above are highly relevant to our project's focus on YouTube influencer reviews. These videos, being inherently multimodal, play a crucial role in the dissemination of information and the shaping of public sentiment [16, 9]. Consequently, understanding the affective dimension of influencer reviews is key to comprehending their impact. This understanding is also vital for businesses, for whom analysing how products are portrayed in such videos offers critical insights [1, 2]. Google Gemini's native multimodal capabilities offer a promising avenue for analysing this type of content comprehensively. Starting from the reviewed state-of-the-art, our project proposes a system that exploits the advanced multimodal understanding capabilities of Google's Gemini model. The central innovation of our work lies in the direct and applied use of Gemini for a specific, high-value business use-case: the detailed analysis of YouTube influencer product reviews to provide companies with actionable, data-driven intelligence. What we also do is to perform a complete analysis on YouTube videos, that fuses together all the features that are typical of videos: audio, video and textual transcription.

Our system will analyse the intricate interplay between: verbal content (derived from accurate transcriptions of the influencer's speech), vocal characteristics (analysing aspects such as tone, pitch, intonation, and speech rate, which carry significant emotional information), visual cues (interpreting the influencer's facial expressions, such as smiles, frowns, and surprise), their body language (where visible and relevant), and the visual presentation and demonstration of the product itself. This approach is designed to offer a significantly more seamless, accurate, and holistic assessment of product perception and influencer sentiment than systems that rely on unimodal (e.g. text-only) analysis or less integrated multimodal approaches. For instance, previous work has involved computing the average of different sentiment analysis outcomes, resulting from separate analyses of text, audio, and video [3].

One additional feature of our system is to give back to the client a structured analysis, and not only a numerical score, like the ones traditionally outputted by classical sentiment analysis. At the same time, the ability of GenAI systems to generate text is exploited in order to produce a detailed and immediate textual feedback on the specific product, that is a fast way for the company to understand the perception of its product among reviewers, and thereby among consumers.

4 System Design and Workflow

This section details the system's design and operational workflow. We present the structured, two-phased methodology developed to transform raw YouTube video reviews from selected influencers into actionable business intelligence. This approach prioritises nuanced understanding and systematic analysis, leveraging Google's Gemini AI at critical junctures for both individual content interpretation and subsequent synthesis of collective insights.

4.1 Phase 1: Data Ingestion and Initial Multimodal Analysis

The first phase focuses on systematically identifying relevant video content, performing an in-depth multimodal analysis of each selected video, and structuring the extracted information for persistent storage. This ensures a high-quality, relevant dataset for subsequent synthesis. The key steps in this phase are:

1. **Product Definition and Source Strategy Formulation:** The process commences with the precise definition of target products for analysis (e.g. specific smartphone models, B2B software versions), including essential metadata such as brand, model, generation/version, and release year. Concurrently, a strategy for identifying relevant video sources is formulated. This may involve curating a list of specific content creators or employing broader search parameters across the YouTube platform, with the detailed rationale and methodology for this choice, including its application to different product categories like B2C smartphones or B2B solutions, further elaborated in Section 5.

2. **Automated Video Discovery:** Utilising the YouTube Data API v3, we programmatically search for videos pertinent to the defined target products. The scope and nature of these searches, which includes configurable parameters such as the target number of videos to retrieve per product or per reviewer-product combination, are directly informed by the source selection strategy (as detailed in Section 5). Accordingly, searches can be narrowly focused within the channels of pre-selected reviewers (if applicable) or executed as wider, keyword-driven queries across YouTube to identify relevant content when specific influencer channels are not predefined. This stage automates the initial identification and collection of potentially relevant video content.

3. **AI-Assisted Relevance Filtering:** To ensure that only substantive, in-depth reviews are subjected to full analysis, each discovered video's title and description undergoes a preliminary assessment. The mentioned metadata is sent to Google's Gemini API with a carefully crafted prompt designed to classify the video's nature (e.g. detailed review vs. news update, brief impressions, or event coverage). Only videos identified as suitable reviews proceed to the next stage.

4. **Multimodal Video Analysis via GenAI:** For each relevant video, we leverage Gemini's advanced multimodal capabilities. Instead of traditional methods requiring separate analysis

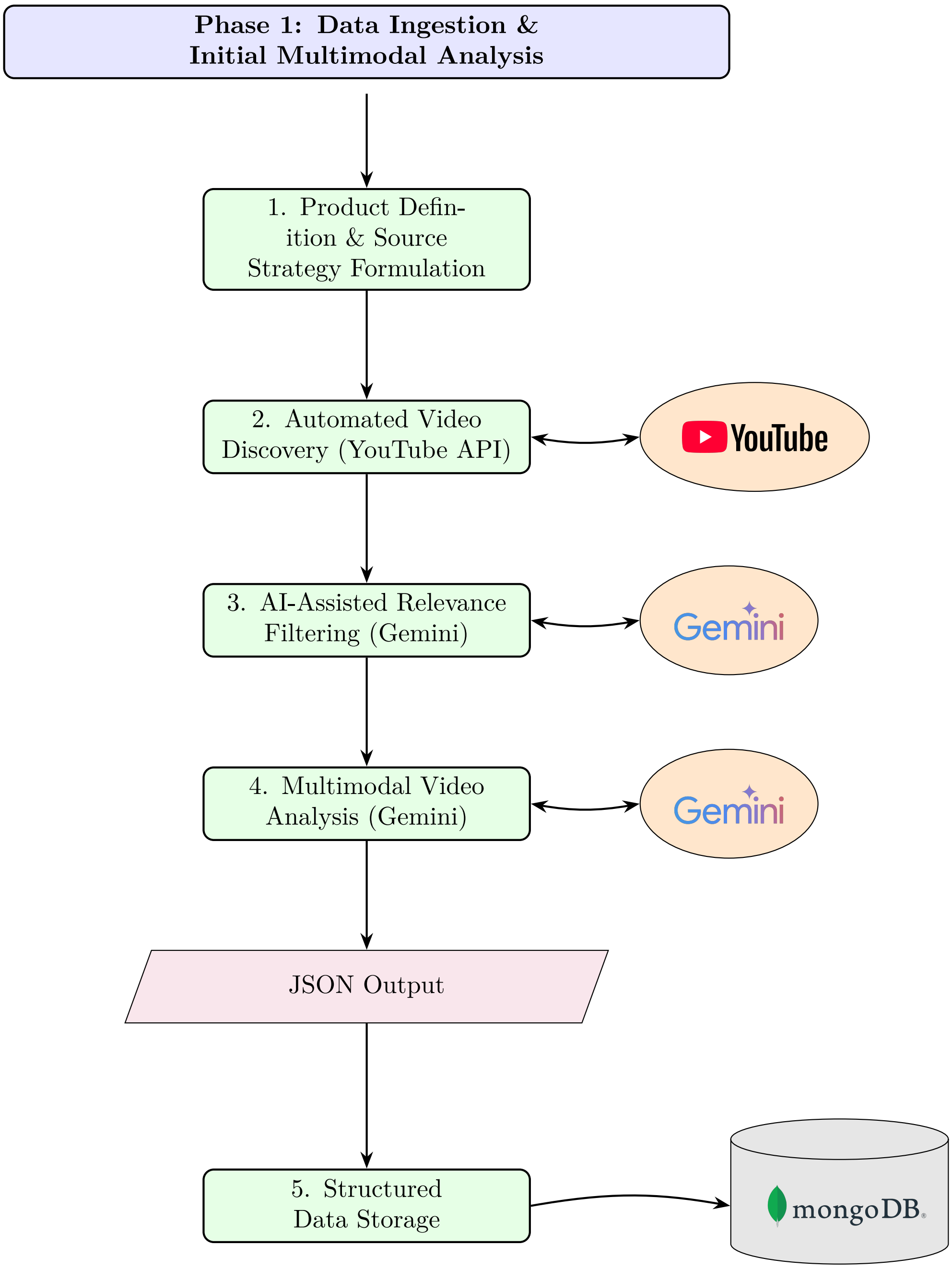
Phase 1: Data Ingestion & Initial Multimodal Analysis
1. Product Definition & Source Strategy Formulation
2. Automated Video Discovery (YouTube API)
YouTube
3. AI-Assisted Relevance Filtering (Gemini)
Gemini
4. Multimodal Video Analysis (Gemini)
Gemini
JSON Output
5. Structured Data Storage
mongoDB

Figure 1: Workflow for Phase 1: Data Ingestion and Initial Multimodal Analysis.

of audio, video, and text, Gemini directly analyses the video content from its YouTube URL. A comprehensive prompt instructs the AI to extract a wide array of information and structure it into a detailed JSON object. This JSON output (requested in English regardless of the video's original language) includes:

- Core video metadata (e.g. title, channel, and the specific product reviewed as identified by the AI).
- An overall sentiment assessment, including an enumerated rating (e.g. positive, neutral, negative), a textual summary, and key positive/negative takeaways.
- A granular feature-by-feature analysis (e.g. camera, battery, design, performance – in the case of smartphones), detailing the sentiment expressed for each feature along with

specific reviewer comments.

  - Insights into the reviewer's perception of pricing and value for money.
  - Notes on any comparisons made to previous product generations or competitor products mentioned within the video.
  - Comments pertaining to brand perception.
  - The target audience for the product, as suggested or implied by the reviewer.
  - A dedicated section on non-verbal cues, which includes an analysis of the reviewer's overall demeanour, notable facial expressions, significant shifts in tone of voice, and prominent gestures, along with their perceived implications on the review's message.

5. **Structured Data Storage:** Each JSON object generated from the analysis of a single video is stored as an individual document in a MongoDB database. These documents are enriched with top-level metadata linking them to our predefined product list (brand, model, generation, release year) and including the video's link and original publication date. Mechanisms are in place to prevent duplicate analyses of the same video-product pair.

4.2 Phase 2: Synthesised Insights and Reporting

With a repository of structured individual review analyses, the second phase focuses on aggregating these individual data points to generate higher-level, synthesised insights. This involves a secondary layer of AI-driven analysis acting upon the data collected in Phase 1.

1. **Targeted Data Loading:** Batches of the stored JSON analyses are retrieved from MongoDB based on specific criteria pertinent to the desired output, such as all analyses for a particular brand (e.g. Apple iPhones across several generations) or all analyses for a set of competing products from a specific year (e.g. 2023 flagship smartphones).

2. **AI-Powered Synthesis of Multiple Reviews:** The selected batches of JSON data, representing multiple individual reviews, are then provided as context to Google's Gemini API. A new, dedicated prompt guides Gemini to act as an analyst, synthesising the collective information to:

   - **Generate Longitudinal Brand/Product Evolution Reports:** For example, to understand how reviewer sentiment and perception of key features for Apple iPhones have evolved over the last n selected generations. The AI is tasked with identifying trends, consistent praises or criticisms, and shifts in opinion over time based on the provided sequence of review data.
   - **Generate Comparative Product Analysis Reports:** For instance, to compare models like the latest iPhone and Samsung flagship based on the synthesised feedback from multiple reviewers. This involves identifying relative strengths and weaknesses and overall reviewer preferences across different product aspects.

3. **Output Generation and Storage:** For each synthesis task (longitudinal or comparative), Gemini is prompted to produce two main outputs:

   - A detailed textual summary in a narrative format, suitable for direct consumption by business stakeholders.
   - A new, structured JSON object that summarises the synthesised findings in a machine-readable format, facilitating further data analysis or integration.

   The generated reports also include metadata, such as the number of source review documents used for each product in the synthesis, providing context on the breadth of input for the generated insights.

This two-phase system design fosters a scalable and highly adaptable approach to extracting deep, nuanced insights from video reviews, moving significantly beyond traditional sentiment

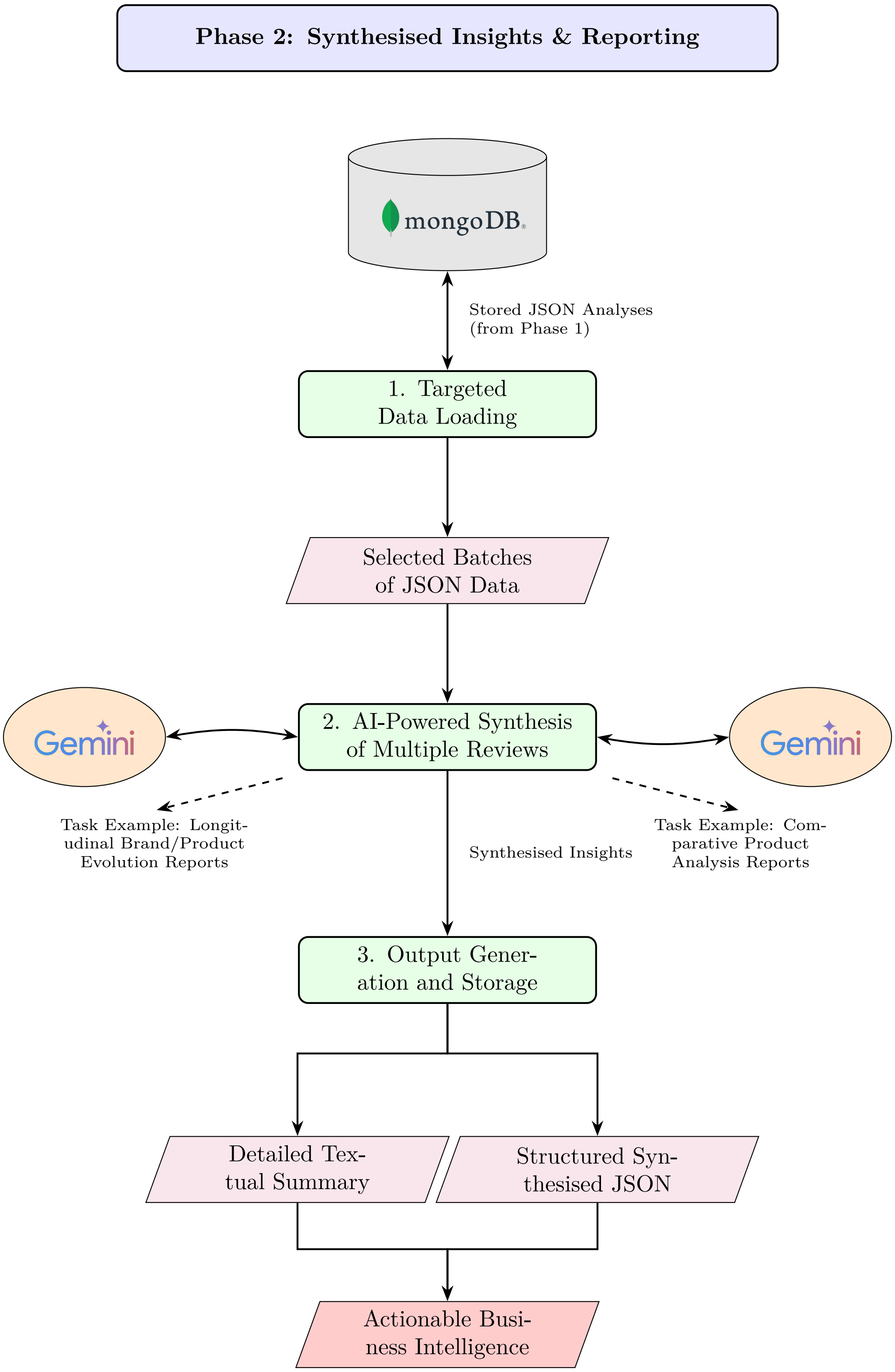
Phase 2: Synthesised Insights & Reporting
mongoDB
Stored JSON Analyses (from Phase 1)
1. Targeted Data Loading
Selected Batches of JSON Data
Gemini
2. AI-Powered Synthesis of Multiple Reviews
Gemini
Task Example: Longitudinal Brand/Product Evolution Reports
Synthesised Insights
Task Example: Comparative Product Analysis Reports
3. Output Generation and Storage
Detailed Textual Summary
Structured Synthesised JSON
Actionable Business Intelligence

Figure 2: Workflow for Phase 2: Synthesised Insights and Reporting.

analysis techniques. A significant advantage of this modular architecture is its inherent extensibility. Incorporating a new product domain into the analysis pipeline primarily involves configuring product-specific parameters (such as its defining characteristics and the key features to be analysed), along with tailoring the AI prompts and the expected structured output from Gemini to suit the new domain. This design significantly reduces the overhead for expanding the system's coverage, allowing for efficient application across diverse product categories.

The structured JSON outputs at both phases are key. While our current work focuses on generating longitudinal and comparative reports, the comprehensively detailed JSON analyses stored from Phase 1 represent a rich data asset. This underlying dataset readily enables a wide range of subsequent, potentially different, ad-hoc queries or future analytical tasks beyond the specific syntheses presented in this project.

5 Source Selection Strategy

The selection of video sources is a critical methodological decision that directly influences the scope, relevance, and ultimate utility of the derived sentiment insights. Given the vastness of YouTube content, a strategic approach to identifying impactful product review videos is paramount. Our project has implemented two distinct source selection strategies, tailored to the differing characteristics of the product categories analysed: consumer-facing (B2C) products like smartphones, and business-to-business (B2B) offerings such as Customer Relationship Management (CRM) software.

5.1 Strategy 1: Curated Influencer Focus (B2C)

For product categories like smartphones, a well-established ecosystem of influential YouTube reviewers often exists. Our primary strategy for these B2C products involves the curation of a list of reputable and impactful reviewers. The rationale for this targeted approach is multifaceted. Firstly, these creators typically command large audiences, and their opinions demonstrably shape consumer perception and purchasing decisions; for businesses, understanding the sentiment projected by these key opinion leaders is crucial for market intelligence. Secondly, many prominent reviewers have built reputations for detailed and relatively consistent analysis, and they often receive products directly from manufacturers for review purposes, ensuring timely and comprehensive coverage. Thirdly, while one might speculate about the influence of sponsorships or affiliate links, the prevailing practice among leading reviewers is an endeavour towards editorial independence; in any event, their extensive viewership ensures their narratives invariably wield significant influence. By analysing content from a diverse set of these top-tier reviewers, we aim to capture a spectrum of prominent opinions rather than relying on a single voice, thereby mitigating the impact of individual biases. Finally, focusing on this known cohort provides a high signal-to-noise ratio, as exploratory searches for smartphone reviews using broader, randomised criteria often yielded videos with significantly lower viewership, less in-depth analysis, or questionable credibility, making them less valuable for understanding mainstream market sentiment.

This curated influencer strategy is extensible to other B2C products where similar, identifiable clusters of influential content creators exist. Companies operating in such markets are often already aware of, or can readily identify through market research, the key voices shaping discussions around their products.

5.2 Strategy 2: Quality-Focused Broader Search (B2B)

For B2B products like CRM software, the "influencer" landscape often differs significantly. A distinct approach was therefore necessary, prioritising a quality-focused broader search rather than pre-defined influencer channels. The key considerations for this strategy include several points. Firstly, dedicated, high-profile "CRM influencers" with massive general appeal are less common compared to B2C tech, so reviews might come from industry consultants, specialised tech channels, or official vendor channels. Secondly, for B2B solutions, the value of a review often lies more in its technical accuracy, feature-by-feature breakdown, and discussion of integration capabilities, rather than sheer viewership numbers; in-depth opinions from knowledgeable sources, even if less "popular", provide critical insights for business decision-makers. Finally, relevant B2B reviews might be found through targeted keyword searches (e.g. "Salesforce vs HubSpot review", "best CRM for small business features"), meaning the focus shifts from known personalities to the quality and depth of the content itself.

This approach involves using the YouTube Data API v3 with carefully constructed search queries to discover relevant videos, followed by rigorous relevance filtering. It is suitable for many B2B products or niche B2C items where a concentrated group of high-reach influencers is not the primary source of valuable review content.

5.3 Relevance Filtering

Irrespective of the initial sourcing strategy (curated list or broader search), a crucial subsequent step in our pipeline is an automated relevance assessment performed by a Gemini model. As detailed in Section 4, this pre-analysis filter examines video titles and descriptions. Its purpose is to confirm that a candidate video is indeed a substantive review suitable for our detailed multimodal sentiment extraction, rather than, for example, a brief news update, an unboxing-only video, or an event recap[1]. This ensures that only the most appropriate content proceeds to the full, resource-intensive analysis, optimising both the quality of the extracted data and the efficiency of API usage.

5.4 Tailoring Source Selection to Specific Business Objectives

While the two primary strategies detailed above address common scenarios for B2C and B2B products, the source selection framework can be further adapted to meet more specific or nuanced business intelligence needs. For instance, a company might require analysis focused on niche product categories, where influence is distributed among smaller, highly specialised content creators not captured by broad B2C or B2B approaches. The focus could also be on hyper-competitive intelligence, involving the deliberate identification and analysis of videos specifically comparing a company's product against a key competitor, regardless of the reviewer's general prominence. Furthermore, analysis might target early trend detection by concentrating on emerging channels or less established voices that might be early indicators of shifting sentiment or new product feature demands. Additionally, there might be a need for regional or demographic specificity, targeting reviewers or content popular within particular geographical markets or among specific consumer demographics.

The underlying video discovery mechanisms can be configured with different keywords, channel lists, or search parameters to accommodate such targeted investigations. This flexibility ensures the system can be tailored to a wide array of strategic analytical goals.

6 Performed Experiments

To validate the proposed methodology and gather data for subsequent synthesis, a series of experiments were conducted focusing on the automated analysis of YouTube product reviews. The primary domain for these experiments was consumer smartphones, given their high review volume and diverse feature sets. Additionally, a smaller-scale experiment was performed on B2B SaaS product reviews (specifically CRM software) to assess the framework's adaptability.

6.1 Smartphone Review Analysis Experiment

The core of our experimental work involved collecting and processing reviews for flagship smartphone models released over the past five years by leading manufacturers, primarily Apple and Samsung.

[1] This reliance on metadata (titles and descriptions) for preliminary screening, as opposed to requiring Gemini to analyse the entirety of each discovered video at this stage, is a pragmatic decision. It is primarily driven by considerations of API usage costs and processing time, which would be substantial for full video analysis on a potentially large number of candidates. Experience has shown that distinguishing substantive reviews from other content types is often relatively straightforward based on these textual cues alone. While manual human curation could also perform this screening, an automated approach, once refined, offers significant efficiency gains. Our current methodology represents an optimised balance developed through iterative testing of various strategies.

6.1.1 Data Corpus Characteristics

- **Reviewer Pool:** Analyses were sourced from a curated list of 11 distinct, internationally recognised YouTube technology reviewers, known for their in-depth smartphone evaluations. These reviewers publish content in various languages, including English, Italian, German, Spanish, French, and Hindi; however, all AI-generated analyses were instructed to produce outputs in English. The list of reviewers includes: AlexiBexi, Andrea Galeazzi, Arun Maini (Mrwhosetheboss), Jojol, Marques Brownlee (MKBHD), Pro Android, Smartworld, Technical Guruji (Gaurav Chaudhary), Technikfaultier, Topes de Gama, and Trakin Tech (Arun Prabhudesai).
- **Product Scope:** The experiment targeted 10 distinct flagship smartphone models.
- **Initial Video Search Scope:** For each of the 11 selected YouTubers and each of the 10 smartphone models, we heuristically decided to search for 5 review videos[2]. This resulted in an initial pool of 550 potential videos. However, not all of these identified videos ultimately met the criteria for successful analysis due to relevance filtering.
- **Video Volume:** A substantial number of review videos were processed for each product to ensure a representative sample of influencer opinion. The breakdown of successfully analysed videos per product is presented in Table 1.

Table 1: Number of Analysed Review Videos per Smartphone Model

Product Brand	Product Model	Videos Analysed
Samsung	Galaxy S21 Ultra	30
Samsung	Galaxy S22 Ultra	26
Samsung	Galaxy S23 Ultra	26
Samsung	Galaxy S24 Ultra	25
Samsung	Galaxy S20 Ultra	23
Apple	iPhone 15 Pro Max	23
Apple	iPhone 12 Pro Max	21
Apple	iPhone 13 Pro Max	20
Apple	iPhone 11 Pro Max	18
Apple	iPhone 14 Pro Max	17
Total Videos Analysed (Smartphones)		**229**

6.1.2 Analysis Process Overview

The Phase 1 analysis process, as comprehensively described in Section 4, was executed for each identified video. This involved automated video discovery via the YouTube Data API v3, AI-assisted relevance filtering, and subsequent multimodal video analysis using Google's Gemini API. The detailed Phase 1 analysis prompt (referenced in Section 7 and available in the project repository) guided Gemini to extract structured JSON outputs. These outputs, rich with sentiment data, feature-specific feedback, and non-verbal cue analysis, were then systematically stored in our MongoDB database. This curated dataset of individual review analyses served as the foundation for the Phase 2 synthesis tasks, which aimed to derive broader insights such as longitudinal brand evolution and comparative product performance reports.

6.2 B2B SaaS (CRM Software) Review Analysis Experiment

To demonstrate the versatility of the analytical framework, a similar process was applied to identify and analyse YouTube reviews focusing on B2B SaaS products, specifically Customer Relationship Management (CRM) software popular among Small to Medium Businesses (SMBs). While the scale of this experiment was smaller, the methodology for video discovery, relevance checking (using a tailored prompt, see `configs/prompts_saas.py` in the repository for an example), Phase

[2] The idea behind the decision is that each youtuber might have more than one video per smartphone, but only up to a couple are likely to be relevant to our analysis (full, in depth reviews).

1 multimodal analysis, and Phase 2 synthesis remained fundamentally the same. This confirmed the adaptability of the prompting strategies and the core AI capabilities to a different product domain with distinct features and evaluation criteria.

Further details on the prompting strategies and examples of the structured outputs for both consumer and B2B analyses are provided in Section 7.

6.3 B2B SaaS (CRM Software) Review Analysis Experiment

To demonstrate the versatility of the analytical framework, a similar process was applied to a curated set of YouTube reviews focusing on B2B SaaS products, specifically Customer Relationship Management (CRM) software popular among Small to Medium Businesses (SMBs). While the scale of this experiment was smaller, the methodology for video discovery, relevance checking (using a tailored prompt, cf. the repository for an example), Phase 1 multimodal analysis, and Phase 2 synthesis remained fundamentally the same.

6.3.1 Data Corpus Characteristics

- **Product Scope:** The experiment targeted 3 prominent CRM software solutions: Salesforce Sales Cloud, HubSpot CRM Suite, and Microsoft Dynamics 365 Sales, chosen for their relevance to SMBs.
- **Video Sourcing and Filtering:** Due to the nature of B2B SaaS products, the pool of suitable, in-depth YouTube reviews is comparatively smaller than for consumer electronics. Consequently, the initial video discovery phase yielded fewer candidates, and the AI-assisted relevance filtering (using a tailored prompt, see `configs/prompts_saas.py` in the repository for an example) played a more critical role in identifying videos that provided substantial evaluative content rather than just tutorials or marketing material. This selection and refinement process was therefore more extensive.
- **Video Volume:** Despite the more challenging sourcing, a sufficient number of high-quality reviews were processed to enable meaningful analysis and synthesis, demonstrating the framework's capability to derive consistent and valuable insights even with a more constrained dataset. The breakdown of successfully analysed videos per CRM product is presented in Table 2.

Table 2: Number of Analysed Review Videos per B2B SaaS (CRM) Product

CRM Product	**Videos Analysed**
Salesforce Sales Cloud	11
HubSpot CRM Suite	10
Microsoft Dynamics 365 Sales	7
Total Videos Analysed (B2B SaaS)	**28**

This experiment confirmed the adaptability of the prompting strategies and the core AI capabilities to a different product domain with distinct features and evaluation criteria. The successful application to B2B SaaS reviews underscores the framework's potential for broader applicability beyond consumer electronics.

Further details on the prompting strategies and examples of the structured outputs for both consumer and B2B analyses are provided in Section 7.

7 Examples of Prompts and System Outputs

To provide a tangible understanding of our system's operational mechanics and the nature of its outputs, this section presents illustrative examples. These showcase key aspects of the prompts provided to Google's Gemini API and representative snippets of the corresponding structured JSON data generated at different stages of our analysis pipeline (see Section 4 for the full workflow).

7.1 Phase 1: Individual Multimodal Video Analysis

This example demonstrates the analysis of a single YouTube product review video. The objective is to extract nuanced sentiment and detailed feature feedback directly from the video content.

7.1.1 Gemini Prompt Core Instructions (Phase 1 – Multimodal Analysis)

The prompt for Phase 1 instructs Gemini to act as an expert multimodal AI analyst. It specifies the video URL and the product being reviewed. Crucially, it demands the entire output be a JSON object, with all textual content within that JSON (summaries, comments, sentiments, etc.) in English, regardless of the video's original language. Key instructions include:

```
Analyse the provided YouTube video (URL: {video_url})
reviewing: '{product_name}'.
ALL extracted information AND your entire response MUST be structured
EXCLUSIVELY in JSON format, AND ALL TEXTUAL CONTENT WITHIN THE JSON
MUST BE IN ENGLISH.

Base your analysis ONLY on the video content, INCLUDING VISUAL AND
AUDITORY CUES. If information is not mentioned or clearly deducible,
use null, empty strings, or empty lists as appropriate.

For 'non_verbal_cues':
- 'overall_reviewer_demeanour': Assess general attitude/energy.
- 'notable_facial_expressions': Up to 3 key moments. Describe context.
- 'tone_of_voice_analysis': Describe shifts indicating emotion.
- 'gestures_and_body_language': Highlight significant gestures.
```

The requested JSON structure includes fields for: video metadata, overall assessment (sentiment, summary, takeaways), feature analysis (per-feature sentiment, comments), pricing and value, comparison context, brand perception, target audience, non verbal cues, and additional elements.

Note: The full operational prompt, detailing the complete JSON schema (`GEMINI_JSON_STRUCTURE_REQUEST`), is available in the project's GitHub repository, in the `codebase/configs/prompts_consumer.py` file.

7.1.2 Example JSON Output Snippets (Phase 1 – MKBHD iPhone 15 Pro Max Review)

The following snippets are extracted from the actual JSON output generated by Gemini for the review of the iPhone 15 Pro Max by Marques Brownlee (MKBHD). The complete JSON document for this analysis can be found in Appendix A.1.

Overall Assessment Snippet:

```
{
    "overall_assessment": {
        "overall_sentiment": "Mixed",
        "summary_review": "The iPhone 15 Pro Max is a refinement of previous models
            , with USB-C being a significant change. While the camera is improved,
            the reviewer notes some potential battery and overheating issues, and
            that the 'Pro' features may not be useful for everyone.",
        "key_positive_takeaways": [
            "USB-C port is a welcome addition.",
            "Video capabilities are still world-class.",
            "Titanium build is lighter."
        ],
        "key_negative_takeaways": [
            "Potential battery and overheating issues.",
            "Some camera features are only slightly improved.",
            "The 'Pro' features may not be useful for everyone."
        ]
    }
}
```

Listing 1: Phase 1 JSON – Overall Assessment Snippet (MKBHD iPhone 15 Pro Max)

Feature Analysis Snippet (USB-C Port):

```
{
    "feature_analysis": [
        // ... other features ...
        {
            "feature_name": "USB-C Port",
            "sentiment": "Positive",
            "specific_comments": "The phone now has a USB-C port, which is a
                welcome change. It supports USB 3.0 data transfer speeds. The
                included cable is USB 2.0, requiring a separate purchase for faster
                 speeds.",
            "key_quote_feature": ""
        }
        // ... other features ...
    ]
}
```

Listing 2: Phase 1 JSON – Feature Analysis Snippet (MKBHD iPhone 15 Pro Max)

Non-Verbal Cues Snippet (Tone of Voice):

```
{
    "non_verbal_cues": {
        // ... other cues ...
        "tone_of_voice_analysis": [
            {
                "segment_description": "When discussing the camera features.",
                "tone_observed": "Enthusiastic",
                "key_tonal_indicators": "Faster pace, upward inflections, emphasis
                    on positive attributes."
            },
            {
                "segment_description": "When discussing the action button.",
                "tone_observed": "Sceptical",
                "key_tonal_indicators": "Slightly sarcastic tone, questioning the
                    usefulness for all users."
            }
        ]
        // ... other cues ...
    }
}
```

Listing 3: Phase 1 JSON – Non-Verbal Cues Snippet (MKBHD iPhone 15 Pro Max)

These snippets illustrate the structured nature and detail level of the individual review analyses, which form the foundation for broader insights.

7.2 Phase 2: Synthesised Insights Generation

This example illustrates how multiple Phase 1 JSON analyses are employed to generate a synthesised report, specifically a longitudinal brand evolution analysis.

7.2.1 Context

The objective here is to understand how sentiment and feature perception for Apple's flagship iPhones have evolved over the last five selected generations, based on the collective feedback from multiple reviewers for each generation.

7.2.2 Input to Synthesis (Conceptual)

The system retrieves all Phase 1 JSON analysis documents from MongoDB pertaining to the Apple iPhone flagships from the iPhone 11 Pro Max (2019) through to the iPhone 15 Pro Max (2023).

7.2.3 Gemini Prompt Core Instructions (Phase 2 – Longitudinal Synthesis)

The prompt for this synthesis task instructs Gemini to act as an AI market analyst. It is provided with a collection of the Phase 1 JSONs. Key instructions include:

```
You are an AI market analyst. You will be provided with multiple
JSON objects, each representing a detailed analysis of an individual
YouTube product review for various generations of Apple iPhone flagships.

Your task is to synthesise this information to generate:
1.  A detailed textual narrative summarising the longitudinal evolution
    of the Apple iPhone flagship line. Focus on overall sentiment trends,
    and the evolution of key features like Camera, Battery Life, Design,
    Performance, Software/Ecosystem, and Price/Value.
2.  A new structured JSON object summarising these synthesised findings,
    including sentiment trends per feature and key generational milestones.

Base your synthesis ONLY on the provided JSON data.
Identify overarching themes, consensus points, and notable shifts in
reviewer opinion over the generations.
The output narrative and JSON should be in English.
```

Note: The actual data payload sent to Gemini would consist of numerous Phase 1 JSON documents.

7.2.4 Example Synthesised Textual Output Snippet (Phase 2 – Apple iPhone Longitudinal)

The following excerpt is from the textual summary generated by Gemini for the longitudinal analysis of Apple iPhone flagships. The full textual summary is extensive and details trends across multiple features and can be found in the project's GitHub repository in the `analyses/` directory.

> The overall sentiment trend towards Apple's iPhone flagships, based on the provided reviews, shows a generally positive trajectory, albeit with fluctuations and increasing scrutiny over value. [. . .] The camera system consistently received praise and saw significant evolution. The iPhone 11 Pro Max marked a turning point with its triple-lens system and improved low-light performance. [. . .] The iPhone 15 Pro Max saw reviewers highlighting improved video quality and skin tone accuracy [. . .]. Design saw a more mixed reception. While the matte finish of the iPhone 11 Pro Max was appreciated, later models faced criticism for their lack of significant design changes. The FineWoven cases introduced with the iPhone 15 Pro Max received overwhelmingly negative feedback [. . .]. Several criticisms recurred across generations. High price was a constant concern [. . .]. Limited base storage [. . .] and slow charging speeds were also frequent complaints.

7.2.5 Example Synthesised JSON Output Snippets (Phase 2 – Apple iPhone Longitudinal)

Below are snippets from the structured JSON output for the same longitudinal analysis. The complete synthesised JSON can be found in Appendix A.2.

Overall Trend Snippet:

```
{
  "analysis_type": "Longitudinal Brand Evolution",
  "brand": "Apple",
  "product_line": "iPhone Flagships (Longitudinal Analysis)",
  "generations_analyzed": [
    [
      "iPhone 15 Pro Max (2023)", // ... other generations ...
      "iPhone 11 Pro Max (2019)"
    ]
  ],
  "overall_sentiment_trend_summary": "The overall sentiment towards Apple’s iPhone
      flagships is generally positive but with increasing scrutiny over value and
      incremental upgrades. Initial enthusiasm for camera and battery improvements
      gives way to more critical assessments of design, software, and competition."
}
```

Listing 4: Phase 2 JSON – Overall Trend Snippet (Apple Longitudinal)

Feature Evolution Snippet (Camera System):

```
{
  "feature_evolution_analysis": [
    {
      "feature_name": "Camera System",
      "evolution_narrative": "The camera system has consistently been a focus,
          evolving from a dual-lens setup in the iPhone 11 Pro Max to more
          sophisticated triple-lens systems... However, Samsung’s zoom capabilities
           are consistently recognized as superior.",
      "sentiment_trend_across_generations": "Strongly Improving",
      "key_generational_milestones": [
        "iPhone 11 Pro Max: Introduction of triple-lens system and Night Mode",
        "iPhone 13 Pro Max: Improved sensors and Cinematic Mode",
        "iPhone 15 Pro Max: Enhanced video quality and skin tone accuracy"
      ]
    }
    // ... other features ...
  ]
}
```

Listing 5: Phase 2 JSON – Feature Evolution Snippet (Apple Longitudinal)

Recurring Themes Snippet:

```
{
  "recurring_positive_themes": [
    "Excellent camera performance, especially in video",
    "Good battery life",
    "Smooth performance and optimized software"
  ],
  "recurring_negative_themes": [
    "High price",
    "Slow charging speed",
    "Incremental upgrades not always worth the cost",
    "Limited base storage"
  ]
}
```

Listing 6: Phase 2 JSON – Recurring Themes Snippet (Apple Longitudinal)

These examples illustrate the progression from individual, detailed multimodal analyses to aggregated, actionable insights, providing a comprehensive overview of product perception and brand evolution as articulated by influential online reviewers.

7.3 Adaptability to B2B Product Analysis: CRM Software Example

As previously discussed, the analytical framework developed for consumer electronics reviews demonstrates notable adaptability to other product categories, among which we have tested it for B2B software and services. To explore this, the methodology was applied to a selection of YouTube reviews comparing popular Customer Relationship Management (CRM) platforms aimed at Small to Medium Businesses (SMBs).

The core process remained consistent: relevant review videos were identified, and a Phase 1 multimodal analysis was performed using Gemini, albeit with a prompt tailored to extract features pertinent to CRM software (e.g. ease of use, contact management, deal/pipeline management, email marketing, reporting, customisation, pricing tiers, and customer support).

Subsequently, these individual CRM review analyses were synthesised in Phase 2 to generate a comparative report. An example of the textual summary output from such a comparative analysis, titled 'Comparative SaaS Analysis Report: SMB CRM Shootout (Test)', which compares Salesforce Sales Cloud, HubSpot CRM Suite, and Microsoft Dynamics 365 Sales, is provided in Appendix A.3. This demonstrates the system's capacity to derive high-level competitive insights from influencer reviews in a distinct B2B context, mirroring the approach used for smartphone analysis. The underlying principles of detailed feature extraction, sentiment assessment, and comparative synthesis prove robust across these differing product domains.

8 Evaluation of Gemini-Generated Video Analyses

The increasing sophistication of MLLMs, in analysing complex video content necessitates robust evaluation strategies. While large-scale benchmarks offer quantitative metrics for aspects like emotion classification [7] or overall model performance on specific datasets [8], understanding the nice amount of different shades of AI interpretation for tasks such as sentiment analysis and conceptual understanding from videos often benefits from direct comparison with human judgment [4, 15, 12]. For instance, Galperin et al. [4] evaluated ChatGPT's sentiment analysis of YouTube video transcripts by measuring concordance with two independent human reviewers, using categories like neutral, positive, negative, or mixed. Liu et al. [15] explored an MLLM's ability to interpret abstract concepts in YouTube Shorts by assessing alignment with human understanding and conducting qualitative analyses of the AI's explanations, advocating for human-centered evaluation methods in multimodal contexts. Furthermore, López Camuñas et al. [12] investigated the alignment between human-annotated labels (for sentiment and facial expressions) and automatic model outputs from video interviews, highlighting challenges such as limited agreement and the importance of considering individual and cultural variability. These studies underscore the value of comparing AI-generated analyses to human annotations, even on smaller, focused subsets, to gain crucial qualitative insights into the AI's performance characteristics, its ability to capture subtleties, and its potential limitations. This principle directly informs the methodology adopted in this section for a qualitative validation.

While the core of this project relies on the automated analysis of YouTube product reviews using Google's Gemini API for scalability, a qualitative validation of the AI's output is essential. This evaluation aims to understand the AI's proficiency in accurately extracting both explicit information and more nuanced elements, such as reviewer sentiment and non-verbal cues. Due to project constraints and team size, this manual validation was performed on a limited, illustrative subset of the ingested video data.

8.1 Quantitative Agreement Analysis: Performed Experiments

To assess the quality of Gemini's video analyses, we opted for a comparative approach, evaluating its outputs against those generated by human annotators. For this purpose, two specific YouTube product review videos were selected: one for the `iPhone 16 Pro Max` and one for the `PRS Silver Sky` electric guitar. The choice of videos in Italian was made to facilitate the annotation process for our Italian-speaking annotators. Furthermore, the video topics were not strictly confined to those analysed in the broader project. This decision was twofold: firstly, to better align with the available annotators' expertise, and secondly, to evaluate Gemini's general video analysis performance and its alignment with human interpretation across a wider range of content, rather than solely within the project's primary domain. The overarching aim here is to understand if Gemini's analyses generally concur with human judgement.

8.1.1 Manual Annotation Methodology and Data Collection

The annotations were performed by a panel of five human annotators (designated Annotator 1 to Annotator 5 for anonymity) and, independently, by the Gemini AI model. All annotators, human and AI alike (when prompted for structured output), utilised identical structured annotation modules (see Appendix B). These modules were designed to capture specific aspects of the video reviews in a categorisable format, facilitating quantitative comparison.

While the main project leverages Gemini to output detailed textual analyses in JSON format, this validation exercise required a more structured, categorical approach to enable numerical quantification of agreement. The use of predefined categories and scales is crucial for transforming qualitative judgements (such as sentiment) into numerical data suitable for statistical analysis (e.g. inter-rater reliability). It is important to note that the rich textual output generated by Gemini in JSON format in the main part of this project can be readily mapped to such categorical or numerical scales, either through specific prompting strategies or subsequent automated processing, including by querying Gemini itself to perform the categorisation based on its own detailed textual analysis.

The data collected through these modules includes:

- Binary judgements (Yes/No, or *Sì/No* in the original annotation forms[3]) on whether specific product features were discussed.
- Ordinal sentiment ratings for each discussed feature. These were captured on a 6-point scale: *Molto Negativo* (Very Negative), *Negativo* (Negative), *Misto* (Mixed), *Neutrale* (Neutral), *Positivo* (Positive), *Molto Positivo* (Very Positive). These categories were numerically coded from 1 (Very Negative) to 6 (Very Positive) respectively for analysis.
- Categorical ratings for overall assessment items (e.g. overall sentiment, quality of audio/visual presentation, reviewer's tone). These ratings are also ordinal, and the specific scales can be found in the annotation modules detailed in Appendix B.

This dataset forms the basis for the inter-rater reliability calculations presented below.

8.2 Metrics for Inter-Rater Reliability

Evaluating the consistency of annotations, whether between human raters (inter-human reliability) or between human raters and an AI model, often involves specific statistical metrics. A widely recognised standard for measuring inter-rater reliability for categorical data is Cohen's Kappa coefficient (κ) [17]. This statistic is valued because it accounts for agreement that might occur purely by chance, offering a more robust measure than simple percentage agreement [18]. The formula for Cohen's Kappa is $\kappa = (P_o - P_e)/(1 - P_e)$, where P_o is the observed proportional agreement and P_e is the expected chance agreement. Weighted versions of Kappa further allow for nuanced assessment of ordinal data by assigning partial credit for disagreements based on their severity [19].

The fundamental principles of Cohen's Kappa, particularly the systematic comparison of categorised judgments and its correction for agreement that might occur by chance, are crucial for a robust assessment of inter-rater reliability [17, 18]. While this establishes Kappa as a standard and valuable statistic, its direct application to all aspects of our experiment's specific data presented some interpretative challenges. For instance, we observed situations where annotators exhibited highly skewed frequency distributions in their use of rating categories (e.g. one annotator predominantly using a single sentiment category for many product features). In such cases, Cohen's Kappa values can be misleadingly low. This occurs because the statistical correction for chance

[3] The annotation forms were provided in Italian to the annotators; *Sì/No* translates to Yes/No respectively.

agreement, inherent to Kappa, can heavily lower the observed agreement, suggesting that much of it was merely coincidental, even when raw agreement levels appear substantial [20][4].

In order to avoid the interpretative complexities of Kappa within our context, and considering our primary goal of obtaining a clear understanding of the actually observed agreement patterns, this study primarily employs two more direct metrics for its main quantitative analysis: Raw Percentage Agreement (RPA) and Weighted Raw Agreement (WRA).

1. Raw Percentage Agreement (RPA): Used to assess agreement on binary nominal judgments (e.g. whether a feature was discussed or not) and for single-instance categorical judgments (e.g. overall video sentiment). RPA provides a direct and easily interpretable measure of exact concordance.

2. Weighted Raw Agreement (WRA): Applied to the ordinal sentiment ratings of product features. This metric calculates the average observed agreement, where disagreements are weighted according to their proximity on the defined ordinal scale, thus capturing nuances in partial agreement. This corresponds to the $P_o(w)$ component of Weighted Kappa, focusing on the observed agreement with partial credit, without the subsequent correction for chance agreement ($P_e(w)$).

This dual approach aims to provide both a direct measure of observed agreement (RPA and WRA) and a qualitative understanding of the AI's performance relative to human annotators.

8.2.1 Computed Agreement Metrics: Definitions

We now provide further details on the metrics computed to assess agreement between annotators (human-human and human-AI).

Raw Percentage Agreement (RPA): RPA measures the simplest form of agreement. For any two raters, it is calculated as the number of times they assigned the exact same category to an item, divided by the total number of items for which *both* raters in that pair provided a valid assessment. If one rater in a pair did not assess an item that the other did (e.g. by leaving it blank or marking it as not applicable where such an option was available and distinct from a substantive category), that item is excluded from the denominator for that specific pairwise calculation. The result is expressed as a percentage. The formula for a specific pair of raters is:

$$\text{RPA} = \left(\frac{\text{Total Number of Agreements}}{\text{Total Number of Items jointly assessed}} \right) \times 100\%$$

RPA was used for:

- **Feature Discussion Agreement (FDA):** This assesses agreement on the binary judgements for whether specific product features were discussed. The annotation modules (see Section I of modules in Appendix B) listed features, and annotators marked each as *Discussa?* (Discussed?) with a *Sì* (Yes) or *No* response. For each pair of raters, the 'number of items jointly assessed' in the RPA formula corresponds to the count of features for which both provided either a 'Sì' or 'No' response.

- **Overall Assessment Agreement (OAA):** This measures agreement on the individual categorical ratings assigned to items in Sections II and III of the annotation modules (e.g. overall sentiment, quality of audio/visual presentation, reviewer's tone). For each such item, it contributes to the RPA calculation for a pair of raters only if both provided a rating for that specific item. The results for multiple such items can then be aggregated or averaged for an overall view of agreement for these sections for the pair.

[4] It is acknowledged that other robust chance-corrected inter-rater reliability measures suitable for multiple raters and various data types exist, such as Fleiss' Kappa [21] for nominal agreement among multiple raters, and Krippendorff's Alpha [22] which offers versatility for different levels of measurement including ordinal data. In our work they were not included for the same reasons discussed about Cohen's Kappa.

Weighted Raw Agreement (WRA) for Ordinal Sentiment: Applied to the ordinal sentiment ratings of product features (Section I). The 6-point sentiment scale (*Molto Negativo*=1 to *Molto Positivo*=6) was used with linear weights $w_{ij} = 1 - \frac{|rank_i - rank_j|}{k-1}$, where $k = 6$ is the number of categories, and $rank_i$, $rank_j$ are the numerical ranks of the categories chosen by rater i and rater j. The WRA for a pair of raters across N_{pair} commonly rated features is:

$$\text{WRA} = \frac{\sum_{\text{item}=1}^{N_{\text{pair}}} w_{ij_{\text{item}}}}{N_{\text{pair}}} \times 100\%$$

This metric gives full credit for exact agreement ($w_{ii} = 1$) and partial credit for "near misses" (e.g. $w_{\text{Positive, Very Positive}} = 1 - \frac{|5-6|}{6-1} = 1 - \frac{1}{5} = 0.8$).

Here, N_{pair} represents the number of features for which *both* raters in a pair: (i) agreed that the feature *was* discussed (i.e. both selected 'Sì' for its discussion status), and (ii) both provided a valid sentiment rating from the 6-point scale.

8.2.2 Statistical Significance of Agreement Differences

To formally assess whether observed differences in agreement scores between the inter-human annotator pairs and the human-AI (Gemini) annotator pairs were statistically meaningful or likely due to sampling variability, a series of non-parametric tests were conducted. The Mann-Whitney U test (also known as the Wilcoxon rank-sum test) was selected for comparing the distributions of the two independent sets of agreement scores (10 inter-human scores vs. 5 human-AI scores for each metric). This test is appropriate given the small sample sizes and makes no assumptions about the normality of the data distributions. For each comparison, the null hypothesis (H_0) stated that there was no difference in the central tendency of agreement scores between the inter-human group and the human-AI group. A significance level of $\alpha = 0.05$ was used for all tests.

8.3 Results of Quantitative Agreement Analysis

The quantitative agreement between the five human annotators and the Gemini AI model was assessed for both the iPhone 16 review video and the PRS Silver Sky guitar review video. The analysis focused on FDA using RPA, FSA using WRA, and OAA also using RPA. For each metric, the set of 10 inter-human agreement scores was compared against the set of 5 human-AI agreement scores using the Mann-Whitney U test, as previously described. The results are summarised below.

iPhone 16 Pro Max Review Video Results: The agreement scores and statistical comparisons for the iPhone 16 Pro Max review video are presented in Table 3.

Table 3: Agreement Results and Statistical Comparison for iPhone 16 Review Video

Agreement Type	Group	N (items)	Mean Agr. (%)	Range Agr. (%)	U-stat	p-value	Sig. Diff? (α=0.05)
FDA (RPA)	Human-Human Human-Gemini	18 18	96.67 96.67	94.44 – 100.00 94.44 – 100.00	25.0	1.0000	No
FSA (WRA)	Human-Human Human-Gemini	≈17-18 ≈17-18	79.75 79.75	66.25 – 92.50 65.00 – 86.25	23.0	0.8536	No
OAA (RPA)	Human-Human Human-Gemini	6 6	63.33 80.00	33.33 – 83.33 50.00 – 100.00	12.0	0.1097	No

For the iPhone 16 review, FDA was exceptionally high and identical for both inter-human and human-Gemini pairs (mean RPA ≈ 96.67 %). The Mann-Whitney U test confirmed no statistically significant difference between the groups for FDA (U=25.0, p=1.0000). For FSA, the mean WRA was also identical for inter-human and human-Gemini pairs (≈ 79.75 %), and the Mann-Whitney U test found no statistically significant difference (U=23.0, p=0.8536). For OAA, while the human-Gemini pairs showed a higher mean RPA (80.00 %) compared to inter-human pairs (63.33 %), this difference was not statistically significant (U=12.0, p=0.1097).

PRS Silver Sky Guitar Review Video Results: Similarly, the agreement scores and statistical comparisons for the PRS Silver Sky guitar review are detailed in Table 4.

Table 4: Agreement Results and Statistical Comparison for PRS Silver Sky Guitar Review Video

Agreement Type	Group	N (items)	Mean Agr. (%)	Range Agr. (%)	U-stat	p-value	Sig. Diff? (α=0.05)
FDA (RPA)	Human-Human Human-Gemini	18 18	100.00 100.00	100.00 – 100.00 100.00 – 100.00	N/A*	N/A*	N/A*
FSA (WRA)	Human-Human Human-Gemini	18 18	88.44 92.00	82.22 – 97.78 82.22 – 98.89	16.5	0.3241	No
OAA (RPA)	Human-Human Human-Gemini	6 6	60.00 73.33	33.33 – 100.00 33.33 – 100.00	17.5	0.3764	No

*Statistical test not performed for FDA due to perfect agreement in both groups.

For the PRS Silver Sky review, FDA was perfect (100% RPA) for all pairs, making statistical comparison unnecessary. For FSA, the human-AI group showed a slightly higher mean WRA (92.00 %, median 93.33 %) than the inter-human group (88.44 %, median 88.89 %), but this difference was not statistically significant (U=16.5, p=0.3241). Similarly, for OAA, the human-AI group had a higher mean RPA (73.33 %, median 83.33 %) compared to the inter-human group (60.00 %, median 50.00 %), but again, the difference was not statistically significant (U=17.5, p=0.3764).

Detailed tables of all pairwise agreement scores are available in the project's GitHub repository.

8.3.1 Summary of Quantitative Findings and Future Considerations

The quantitative agreement analyses presented for both review videos indicate that Google's Gemini model performs comparably to human annotators across several key video analysis tasks: identifying discussed features, assessing feature-specific sentiment, and evaluating overall video characteristics. In all instances, statistical tests (Mann-Whitney U) did not reveal significant differences between human-human agreement levels and human-Gemini agreement levels. For some metrics, particularly Overall Assessment Agreement and Feature Sentiment Agreement on the PRS Silver Sky video, Gemini's alignment with human annotators even showed numerically higher mean agreement scores than the average inter-human agreement, although these differences were not statistically significant with the current sample size.

These findings are encouraging regarding Gemini's potential for automated video analysis. However, it is important to contextualise these results. This validation was conducted on a limited set of two videos and with a small panel of five human annotators. While illustrative, a more formal and comprehensive evaluation would be necessary to draw definitive conclusions about Gemini's generalisability and robustness. Future research should aim to expand this type of validation by:

- Increasing the number and diversity of videos analysed.
- Engaging a larger and potentially more varied pool of human annotators to enhance statistical power and reduce the impact of individual rater biases.
- Exploring agreement on a wider range of nuanced interpretative tasks.

Furthermore, while this study provides a general indication of Gemini's capabilities, any commercial entity or organisation intending to leverage such MLLM-based video analysis for specific business applications should conduct its own targeted validation, essential to ascertain the model's suitability and reliability for their unique operational context and to fine-tune prompting strategies for optimal performance.

9 Assessment of Multimodal Utility in Sentiment Analysis

A core premise of this project is the utilisation of multimodal GenAI to achieve a more nuanced understanding of sentiment in YouTube product reviews than would be possible from text-alone analysis. This section reflects on the perceived utility of incorporating non-verbal cues (visual and auditory elements) alongside textual content, particularly within the context of our primary use cases: B2C and B2B product reviews.

The direct analysis of video content, encompassing visual information about the reviewer's expressions and gestures, as well as auditory information such as tone of voice, theoretically offers a richer dataset for sentiment interpretation. Models like Gemini, capable of processing these multiple modalities simultaneously, are designed to capture subtleties that might be lost or misinterpreted in a purely transcript-based approach. For instance, sarcasm, enthusiasm, hesitation, or scepticism are often conveyed more strongly through vocal inflection and facial expressions than through transcribed words alone. Our Phase 1 analysis prompt specifically instructed Gemini to identify and interpret such non-verbal cues, aiming to enrich the overall sentiment assessment and provide deeper context to the reviewer's explicit statements.

In the context of the smartphone reviews analysed for this project, the contribution of non-verbal cue analysis to the *overall product sentiment rating* (e.g. positive, mixed, negative for the product as a whole) appeared to be more confirmatory than transformative for many of the professional reviewers selected. These reviewers often articulate their overall conclusions quite explicitly in their verbal summaries. In such cases, the non-verbal cues (e.g. a measured tone when delivering a mixed verdict, or enthusiastic gestures when praising a feature) generally aligned with and reinforced the explicitly stated sentiment. The non-verbal analysis, therefore, often served to increase confidence in the AI's interpretation of the dominant sentiment rather than fundamentally altering it from what might be gleaned from a high-quality transcript. For example, a reviewer might state a feature is "good", and their enthusiastic tone and positive facial expression would confirm this positive sentiment, whereas a flat tone might have suggested a less impactful "good".

However, the value of multimodal analysis becomes more apparent when considering specific feature assessments or identifying subtle shifts in opinion that are not overtly stated. A reviewer might verbally describe a feature with neutral language, but a fleeting expression of frustration, a sceptical tone, or a dismissive gesture could indicate underlying reservations missed by text-only analysis. Our framework attempted to capture these through the `non_verbal_cues` section in the Phase 1 JSON output. While a systematic quantitative comparison of multimodal versus text-only AI analysis was beyond the scope of this initial project phase, the qualitative review of Gemini's multimodal outputs suggests it does pick up on these cues, as seen in the MKBHD case study (Section 7.1.2) where for example a "sceptical tone" was noted for the Action Button discussion.

While the explicit verbal content from skilled professional reviewers often clearly conveys their overall product assessment, the strategic importance of analysing non-verbal cues grows significantly when the analytical goals extend beyond general sentiment or explore more nuanced interpretations. For a basic understanding of a product's reception among top-tier communicators, transcribed text might offer a primary overview. However, for deeper insights, or when examining different types of video content, multimodal analysis becomes indispensable, unlocking potential far beyond immediate business needs. Consider, for example, detecting sarcasm or irony, which are frequently signalled through vocal tone and facial expression rather than words alone; multimodal AI is better equipped to catch these subtleties. This capability for deeper understanding also extends to assessing the genuine enthusiasm behind an influencer's statements, a crucial factor when evaluating sponsored content or reviews from emerging voices. Indeed, the power to interpret these non-verbal signals has implications across various domains: from understanding the implicit messaging in political discourse and enhancing educational feedback by analysing student presentations, to offering new perspectives in areas like mental health through the study of video journals (an application explored, for instance, in the context of depression through YouTube Shorts by Liu et al. [15]). Even within business contexts like brand safety monitoring, subtle non-verbal cues of discomfort associated with a brand might offer early warnings, while unvoiced user frustrations with a product could be revealed through micro-expressions. For less formally structured content, such as live streams or informal vlogs, where the verbal script is less polished, non-verbal signals inherently carry even more weight in conveying true sentiment. In all these varied contexts, the ability of a multimodal AI to process information beyond the transcript is not merely an add-on but a crucial component for a more complete and accurate understanding of the communicated message.

Therefore, while for the specific professional tech reviews analysed in our primary use case, the non-verbal elements often corroborated the explicit verbal sentiment for the overall product, their utility in detecting subtler opinions, verifying authenticity, and enabling a broader range of video content analysis remains significant. Future work could involve a direct comparative study by processing video transcripts alone through a similar AI pipeline and contrasting those outputs with the current multimodal results to more quantitatively assess the uplift provided by visual and auditory analysis across different types of review content and business objectives.

10 Broader Applications and Future Directions

Our proof-of-concept system successfully shows how multimodal GenAI can offer a nuanced analysis of influencer product reviews. Its methodology and framework meet current goals and show considerable promise for wider business intelligence applications and a deeper, broader understanding of video content.

The system's flexible architecture, particularly its adaptable GenAI prompting and the definition of analytical targets, allows for customisation beyond the source selection strategies mentioned in Section 5.4. For instance, longitudinal tracking offers a compelling application: the system could monitor initial reactions to pre-launch units (often seeded to influencers) and then track sentiment as comprehensive reviews and long-term usage reports appear. This can reveal shifts in public and expert perception, quickly identifying emerging issues or unexpected successes vital for agile product management. Alternatively, employing more open-ended prompting with models like Gemini could uncover *unexpectedly* prominent features, user-devised use cases, or recurrent, unaddressed pain points. This offers a valuable, unsolicited feedback channel directly influencing product development and innovation.

Looking further, the core multimodal analysis capabilities are not limited to YouTube. With appropriate API access and content suitability checks, the system could, in principle, dissect video from other influential platforms. This would require adapting data ingestion and relevance filtering but would significantly broaden market insights. Moreover, the nuanced sentiment analysis, especially interpreting non-verbal cues, could be adapted for brand safety and crisis monitoring. The system could be configured to detect subtle negative sentiment or problematic associations with a brand, even if unstated, acting as an early warning for potential reputational risks. The strategic value of these insights can be amplified by integrating the structured sentiment data with a company's internal metrics, such as sales figures or customer support tickets.

The maturation of AI-driven video analysis, as shown by this project, also points to wider societal and technological transformations. As these models become more adept at understanding and generating content, their ability to automatically extract detailed information, identify patterns, and summarise complex events from video will be invaluable across many fields. Applications could range from enhanced media archiving in media studies and objective analysis of public appearances in political science, to personalised learning feedback from educational recordings and tracking behavioural indicators in therapeutic video journals for mental health.

However, this increasing sophistication introduces a critical challenge: the spread of AI-generated or manipulated video content. A crucial future direction for AI video analysis will therefore be to develop and integrate robust mechanisms for detecting AI-generated or deepfake content. Ensuring the authenticity of source material is paramount for maintaining analytical integrity.

11 Conclusions

The analysis of sentiment and emotion embedded within digital content, particularly from complex multimodal sources such as YouTube videos, represents a dynamic and rapidly advancing frontier in artificial intelligence. The existing body of research provides a robust foundation in unimodal and, increasingly, sophisticated multimodal analysis techniques. However, the direct and practical application of cutting-edge GenAI models like Google's Gemini for a holistic, end-to-end analysis of influencer reviews specifically tailored for business intelligence presents a significant and new opportunity.

Our project strategically addresses this opportunity. By developing a system capable of dissecting the multifaceted layers of communication (verbal, visual, and vocal cues) present in YouTube influencer content, we aim to provide businesses with a deeper, more accurate, and actionable understanding of how their products are perceived. While this constitutes a significant business application, the underlying multimodal analysis technology possesses a versatility that extends far beyond this initial use case. We envision its adaptation for diverse fields and purposes, including applications in education, political science, and healthcare, offering novel methods to extract understanding from rich video data. This work, therefore, not only contributes to the field of applied AI and offers a tangible solution for companies navigating the complexities of the modern digital marketplace but also highlights the broader potential of such systems to empower various domains with nuanced insights previously challenging to obtain at scale.

References

[1] Z. Wang, P. Gao, and X. Chu, "Sentiment analysis from Customer-generated online videos on product review using topic modeling and Multi-attention BLSTM," *Advanced Engineering Informatics*, vol. 52, p. 101588, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S147403462200060X?via%3Dihub

[2] A. Usmani, S. H. Alsmadi, M. J. K. Jaleel, J. Breslin, and E. Curry, "MuSe-CarASTE: A comprehensive dataset for aspect sentiment triplet extraction in automotive review videos," *Expert Systems with Applications*, vol. 262, p. 125695, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417424025624

[3] K. Vasantha, P. Sridevivijaymala, V. Shete, R. C N, and V. Sai Krithik Pa, "Dynamic Fusion of Text, Video and Audio models for Sentiment Analysis," *Procedia Computer Science*, vol. 215, pp. 211–219, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922020956

[4] S. Galperin, L. Wiener, O. Chafitz, S. Bittman, and A. F. Oladipo, "A sentiment analysis of YouTube videos from donor-conceived people, utilizing artificial intelligence (ChatGPT)," *Reproductive BioMedicine Online*, 2025. [Online]. Available: https://www.rbmojournal.com/article/S1472-6483(25)00114-2/abstract

[5] M. I. Radaideh, O. H. Kwon, and M. I. Radaideh, "Fairness and social bias quantification in Large Language Models for sentiment analysis," *Knowledge-Based Systems*, vol. 319, p. 113569, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070512500615X

[6] C. Wei, S. Noh, and H.-C. H. Chang, "Faces Speak Louder Than Words: Emotions Versus Textual Sentiment in the 2024 USA Presidential Election," *arXiv preprint arXiv:2412.18031*, 2025, v2 [cs.SI] 25 Mar 2025. [Online]. Available: https://arxiv.org/abs/2412.18031

[7] A. Fernández and S. Awinat, "Multimodal Sentiment Analysis based on Video and Audio Inputs," *Procedia Computer Science*, vol. 251, pp. 41–48, 2024. [Online]. Available: https://arxiv.org/abs/2412.09317

[8] H. Shi, "A short video sentiment analysis model based on multimodal feature fusion," *Systems and Soft Computing*, vol. 6, p. 200148, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772941924000772?via%3Dihub

[9] S. Hofmann, C. Sommermann, M. Kraus, P. Zschech, and J. Rosenberger, "Hate Speech and Sentiment of YouTube Video Comments From Public and Private Sources Covering the Israel-Palestine Conflict," *arXiv preprint arXiv:2503.10648*, 2025, v1 [cs.CL] 3 Mar 2025. [Online]. Available: https://arxiv.org/abs/2503.10648

[10] K. Hegde and H. Jayalath, "Emotions in the Loop: A Survey of Affective Computing for Emotional Support," *arXiv preprint arXiv:2505.01542*, 2025, v1 [cs.HC] 2 May 2025. [Online]. Available: https://arxiv.org/abs/2505.01542

[11] M. J. Sekiavandi, L. Dixen, J. Fimland, S. K. Desu, A.-B. Zserai, Y. S. Lee, M. Barrett, and P. Burreli, "Advancing Face-to-Face Emotion Communication: A Multimodal Dataset (AFFEC)," *arXiv preprint arXiv:2504.18969*, 2025, v2 [cs.HC] 1 May 2025. [Online]. Available: https://arxiv.org/abs/2504.18969

[12] J. Lopez Camuñas, C. Bustos, Y. Zhu, R. Ros, and A. Lapedriza, "Experimenting with Affective Computing Models in Video Interviews with Spanish-speaking Older Adults," *arXiv preprint arXiv:2501.16870*, 2025, v1 [cs.CV] 28 Jan 2025. [Online]. Available: https://arxiv.org/abs/2501.16870

[13] H. Zhang, Z. Li, Y. Zhu, H. Xu, P. Wang, H. Zhu, J. Zhou, and J. Zhang, "Can Large Language Models Help Multimodal Language Analysis? MMLA: A Comprehensive Benchmark," *arXiv preprint arXiv:2504.16427*, 2025, v2 [cs.CL] 24 Apr 2025. [Online]. Available: https://arxiv.org/abs/2504.16427

[14] D. Li, B. Xing, X. Liu, B. Xia, B. Wen, and H. Kälviäinen, "DEEMO: De-identity Multimodal Emotion Recognition and Reasoning," *arXiv preprint arXiv:2504.19549*, 2025, v1 [cs.CV] 28 Apr 2025. [Online]. Available: https://arxiv.org/abs/2504.19549

[15] J. L. Liu, Y. Su, and P. Seth, "Can Large Language Models Grasp Concepts in Visual Content? A Case Study on YouTube Shorts about Depression," *arXiv preprint arXiv:2503.05109*, 2025, v1 [cs.HC] 7 Mar 2025. [Online]. Available: https://arxiv.org/abs/2503.05109

[16] V. Su and N. Thakur, "COVID-19 on YouTube: A Data-Driven Analysis of Sentiment, Toxicity, and Content Recommendations," 2024. [Online]. Available: https://arxiv.org/abs/2412.17180

[17] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[18] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[19] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[20] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: use, interpretation, and sample size requirements," *Physical therapy*, vol. 85, no. 3, pp. 257–268, 2005.

[21] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions.* john wiley & sons, 2013.

[22] K. Krippendorff, "Reliability in content analysis: Some common misconceptions and recommendations," *Human communication research*, vol. 30, no. 3, pp. 411–433, 2004.

Appendices

A Detailed AI Analysis Outputs

A.1 Phase 1: Full JSON Output for MKBHD iPhone 15 Pro Max Review

The following is the complete JSON output generated by Gemini for the review of the iPhone 15 Pro Max by Marques Brownlee (MKBHD) as referenced in Section 7.1.2.

```
{
    "_id": {
        "$oid": "681f2c589b5c1ae46d73da68"
    },
    "product_config_name": "iPhone 15 Pro Max",
    "product_brand": "Apple",
    "product_generation": "15 Pro Max",
    "product_release_year": 2023,
    "video_id": "cBpGq-vDr2Y",
    "video_url": "https://www.youtube.com/watch?v=cBpGq-vDr2Y",
    "video_title_yt": "iPhone 15 Pro Review: The Good, The Bad, & The Ugly!",
    "video_published_at": {
        "$date": "2023-09-28T22:01:44Z"
    },
    "reviewer_channel_id": "UCBJycsmduvYEL83R_U4JriQ",
    "reviewer_name": "Marques Brownlee (MKBHD)",
    "analysis_timestamp": {
        "$date": "2025-05-10T10:37:12.787Z"
    },
    "gemini_analysis": {
        "video_metadata": {
            "video_url": "https://www.youtube.com/watch?v=cBpGq-vDr2Y",
            "video_title": "iPhone 15 Pro Review: The Good, The Bad, & The Ugly!",
            "channel_name": "Marques Brownlee",
            "product_reviewed": "iPhone 15 Pro Max"
        },
        "overall_assessment": {
            "overall_sentiment": "Mixed",
            "sentiment_score_numeric": null,
            "summary_review": "The iPhone 15 Pro Max is a refinement of previous
                models, with USB-C being a significant change. While the camera is
                improved, the reviewer notes some potential battery and overheating
                 issues, and that the ’Pro’ features may not be useful for everyone
                .",
            "key_positive_takeaways": [
                "USB-C port is a welcome addition.",
                "Video capabilities are still world-class.",
                "Titanium build is lighter."
            ],
            "key_negative_takeaways": [
                "Potential battery and overheating issues.",
                "Some camera features are only slightly improved.",
                "The ’Pro’ features may not be useful for everyone."
            ]
        },
        "feature_analysis": [
            {
                "feature_name": "Build",
                "sentiment": "Positive",
                "specific_comments": "Titanium build is lighter and corners are
                    softened. Bezels are slightly thinner. New colors are available
                    .",
                "key_quote_feature": ""
            },
            {
                "feature_name": "Chip (A17 Pro)",
                "sentiment": "Positive",
                "specific_comments": "The A17 Pro is a 3nm chip that is 10-20% more
```

```
                     powerful than the previous generation. It's approaching M1
                    chip levels of performance.",
                "key_quote_feature": ""
            },
            {
                "feature_name": "Camera",
                "sentiment": "Positive",
                "specific_comments": "New 48MP main camera with improved ultra-wide
                     and macro capabilities. The phone shoots 24MP images by
                    default. There is also ProRES video recording.",
                "key_quote_feature": ""
            },
            {
                "feature_name": "Action Button",
                "sentiment": "Neutral",
                "specific_comments": "Customizable button that replaces the mute
                    switch. Can be mapped to various functions, including launching
                     apps and Siri shortcuts. The reviewer notes that it may be
                    overrated or underrated depending on the user.",
                "key_quote_feature": ""
            },
            {
                "feature_name": "USB-C Port",
                "sentiment": "Positive",
                "specific_comments": "The phone now has a USB-C port, which is a
                    welcome change. It supports USB 3.0 data transfer speeds. The
                    included cable is USB 2.0, requiring a separate purchase for
                    faster speeds.",
                "key_quote_feature": ""
            },
            {
                "feature_name": "Battery Life",
                "sentiment": "Mixed",
                "specific_comments": "Objectively, the phone has a slightly bigger
                    battery. However, the reviewer experienced some bad draining
                    days, and the battery life is likely to even out to be the same
                     as last year.",
                "key_quote_feature": ""
            },
            {
                "feature_name": "Overheating",
                "sentiment": "Negative",
                "specific_comments": "The reviewer notes that there are some
                    overheating issues that have been reported.",
                "key_quote_feature": ""
            }
        ],
        "pricing_and_value": {
            "price_mention": false,
            "price_currency": "",
            "price_amount": null,
            "price_sentiment": "Not Mentioned",
            "value_for_money_assessment": "The reviewer does not explicitly state
                if the phone is worth the money, but implies that it is only worth
                it if the new features are important to you."
        },
        "comparison_context": {
            "vs_previous_generation": {
                "mentioned": true,
                "previous_product_name": "iPhone 14 Pro",
                "key_differences_highlighted": "USB-C, faster chip, brighter screen
                    , thinner bezels, better cameras, titanium build, action button
                    ",
                "overall_comparison_sentiment": "Improvement"
            },
            "vs_competitors": [
                {
                    "competitor_name": "Samsung",
                    "comparison_points": "Telephoto lens reach",
                    "outcome": "Samsung has a longer telephoto lens reach with its
```

```
                        10x optical zoom."
                }
            ]
        },
        "brand_perception": {
            "brand_sentiment": "Neutral",
            "brand_related_comments": "The reviewer mentions that Apple is being
                forced to use USB-C by the EU."
        },
        "target_audience": {
            "suggested_by_reviewer": "The reviewer suggests the phone is best for
                those who need the Pro features, especially the improved video
                capabilities."
        },
        "non_verbal_cues": {
            "overall_reviewer_demeanour": "Measured",
            "demeanour_justification": "The reviewer maintains a calm and
                professional tone throughout the video, presenting information in a
                 balanced manner.",
            "notable_facial_expressions": [
                {
                    "expression_type": "Raised Eyebrows",
                    "context_description": "When mentioning the ’RAW Max’ button
                        name.",
                    "perceived_implication": "Mild amusement or skepticism towards
                        Apple’s naming choice."
                },
                {
                    "expression_type": "Smile",
                    "context_description": "When describing the USB-C port.",
                    "perceived_implication": "Genuine approval of the change."
                },
                {
                    "expression_type": "Neutral",
                    "context_description": "When describing the battery life.",
                    "perceived_implication": "The reviewer is being objective and
                        not trying to oversell the battery life."
                }
            ],
            "tone_of_voice_analysis": [
                {
                    "segment_description": "When discussing the camera features.",
                    "tone_observed": "Enthusiastic",
                    "key_tonal_indicators": "Faster pace, upward inflections,
                        emphasis on positive attributes."
                },
                {
                    "segment_description": "When discussing the battery and
                        overheating issues.",
                    "tone_observed": "Neutral",
                    "key_tonal_indicators": "Calm, even tone, objective
                        presentation of facts."
                },
                {
                    "segment_description": "When discussing the action button.",
                    "tone_observed": "Sceptical",
                    "key_tonal_indicators": "Slightly sarcastic tone, questioning
                        the usefulness for all users."
                }
            ],
            "gestures_and_body_language": [
                {
                    "gesture_description": "Using hands to frame the phone.",
                    "context_description": "When discussing the build and design
                        changes.",
                    "perceived_implication": "Emphasizing the physical attributes
                        of the phone."
                },
                {
                    "gesture_description": "Shrugging shoulders.",
                    "context_description": "When describing the camera's
                        performance.",
                    "perceived_implication": "Uncertainty or lack of strong
                        conviction about the improvements."
                },
                {
                    "gesture_description": "Pointing at the phone.",
                    "context_description": "When highlighting specific features of
                        the phone.",
                    "perceived_implication": "Directing attention to key aspects of
                         the device."
                }
            ]
        },
        "additional_elements": {
            "key_quote_overall_positive": "It's the same phone as last year with
                USB.",
            "key_quote_overall_negative": "There almost seems to be no rhyme or
                reason why it does this randomly once in a while.",
            "notable_mentions": [
                "The reviewer mentions the 8Sleep mattress cover as a product that
                    has improved his sleep.",
                "The reviewer mentions that the 4x zoom on the iPhone 15 Pro may
                    actually be worse than the 5x zoom."
            ]
        }
    }
}
```

Listing 7: Complete Phase 1 JSON – MKBHD iPhone 15 Pro Max

A.2 Phase 2: Full Synthesised JSON Output for Apple iPhone Longitudinal Analysis

The following is the complete synthesised JSON output generated by Gemini for the longitudinal analysis of Apple iPhone flagships (2019–2023), as referenced in Section 7.2.5.

```
{
  "analysis_type": "Longitudinal Brand Evolution",
  "brand": "Apple",
  "product_line": "iPhone Flagships (Longitudinal Analysis)",
  "generations_analyzed": [
    [
      "iPhone 15 Pro Max (2023)",
      "iPhone 14 Pro Max (2022)",
      "iPhone 13 Pro Max (2021)",
      "iPhone 12 Pro Max (2020)",
      "iPhone 11 Pro Max (2019)"
    ]
  ],
  "overall_sentiment_trend_summary": "The overall sentiment towards Apple's iPhone
      flagships is generally positive but with increasing scrutiny over value and
      incremental upgrades. Initial enthusiasm for camera and battery improvements
      gives way to more critical assessments of design, software, and competition."
      ,
  "feature_evolution_analysis": [
    {
      "feature_name": "Camera System",
      "evolution_narrative": "The camera system has consistently been a focus,
          evolving from a dual-lens setup in the iPhone 11 Pro Max to more
          sophisticated triple-lens systems with improved sensors, zoom
          capabilities, and software processing. The iPhone 15 Pro Max sees a focus
           on video quality and skin tone accuracy. However, Samsung's zoom
          capabilities are consistently recognized as superior.",
      "sentiment_trend_across_generations": "Strongly Improving",
      "key_generational_milestones": [
        "iPhone 11 Pro Max: Introduction of triple-lens system and Night Mode",
        "iPhone 13 Pro Max: Improved sensors and Cinematic Mode",
        "iPhone 15 Pro Max: Enhanced video quality and skin tone accuracy"
      ]
    },
    {
      "feature_name": "Battery Life",
      "evolution_narrative": "Battery life saw a significant leap with the iPhone
          11 Pro Max and remained a strength in subsequent models. However, the
          iPhone 15 Pro’s battery life was considered underwhelming, highlighting
          the fluctuating nature of this feature. Optimization of software and
          hardware plays a key role in perceived battery performance.",
      "sentiment_trend_across_generations": "Moderately Improving",
      "key_generational_milestones": [
        "iPhone 11 Pro Max: Major leap in battery life",
        "iPhone 15 Pro: Underwhelming battery life compared to expectations"
      ]
    },
    {
      "feature_name": "Performance",
      "evolution_narrative": "Performance has consistently been excellent across
          all generations due to Apple’s A-series chips. While each new chip offers
           a performance boost, reviewers often note that the real-world
          differences are minimal for everyday tasks.",
      "sentiment_trend_across_generations": "Consistent Positive",
      "key_generational_milestones": [
        "All generations: Consistent praise for A-series chip performance"
      ]
    },
    {
      "feature_name": "Design",
      "evolution_narrative": "Design has been a more polarizing aspect. The matte
          finish of the iPhone 11 Pro Max was appreciated, but later models faced
          criticism for their lack of significant design changes. The titanium
          design of the iPhone 15 Pro received mixed reviews, and the FineWoven
          cases were universally panned for their poor durability.",
      "sentiment_trend_across_generations": "Mixed/Fluctuating",
      "key_generational_milestones": [
        "iPhone 11 Pro Max: Introduction of matte finish",
        "iPhone 15 Pro Max: Titanium design and FineWoven cases"
      ]
    },
    {
      "feature_name": "Software/Ecosystem",
      "evolution_narrative": "The iOS ecosystem is consistently praised for its
          seamless integration and long-term software support. However, reviewers
          also note the limitations in customization compared to Android and the
          occasional software bugs.",
      "sentiment_trend_across_generations": "Consistent Positive",
      "key_generational_milestones": [
        "All generations: Consistent praise for ecosystem integration and long-term
             support"
      ]
    },
    {
      "feature_name": "Price/Value",
      "evolution_narrative": "High price has been a recurring criticism across all
          generations. Reviewers often question whether the incremental upgrades
          justify the cost, especially compared to previous models or Android
          alternatives. Value for money is a major consideration for potential
          buyers.",
      "sentiment_trend_across_generations": "Consistent Negative",
      "key_generational_milestones": [
        "All generations: Consistent criticism of high price point"
      ]
    },
    {
      "feature_name": "Charging Speed",
      "evolution_narrative": "Charging speed has consistently been a point of
          criticism, with reviewers noting that Apple lags behind competitors in
          this area. The lack of a charger in the box is also a recurring complaint
          .",
      "sentiment_trend_across_generations": "Consistent Negative",
      "key_generational_milestones": [
        "All generations: Slow charging speeds compared to competitors"
      ]
    }
  ],
  "recurring_positive_themes": [
    "Excellent camera performance, especially in video",
    "Good battery life",
    "Smooth performance and optimized software"
  ],
  "recurring_negative_themes": [
    "High price",
    "Slow charging speed",
    "Incremental upgrades not always worth the cost",
    "Limited base storage"
  ],
  "significant_opinion_shifts_observed": [
    {
      "generation_product_name": "iPhone 15 Pro Max",
      "shift_description": "Increased focus on sustainability and environmental
          impact in reviews.",
      "supporting_evidence_summary": "Reviewers mention Apple's focus on carbon
          neutrality and the use of recycled materials."
    },
    {
      "generation_product_name": "iPhone 14 Pro Max",
      "shift_description": "The Dynamic Island, initially seen as innovative, later
           faced scrutiny for its limited functionality and screen real estate
          usage.",
      "supporting_evidence_summary": "Early reviews praised the Dynamic Island,
          while later reviews questioned its usefulness and found it distracting."
    },
    {
      "generation_product_name": "iPhone 12 Pro Max",
      "shift_description": "The absence of a fingerprint sensor became a
          significant drawback due to the COVID-19 pandemic and mask-wearing.",
      "supporting_evidence_summary": "Reviewers specifically mentioned the
          inconvenience of Face ID when wearing a mask."
    }
  ],
  "overall_evolution_conclusion": "Apple's iPhone Flagships (Longitudinal Analysis)
       have consistently delivered premium experiences with strong camera
      performance and software integration. However, the evolution is marked by
      incremental improvements, increasing scrutiny over value, and the need to
      address criticisms regarding design choices and charging speed to maintain
      its competitive edge."
}
```

Listing 8: Complete Phase 2 Synthesised JSON – Apple iPhone Flagships (Longitudinal)

A.3 Phase 2: Synthesised Textual Output for B2B CRM Comparative Analysis

The following is the synthesised textual summary generated by Gemini for a comparative analysis of selected CRM software platforms, as referenced in Section 7.3. More examples of such analyses can be found in the GitHub repository under the `analyses/` directory.

Comparative SaaS Analysis Report: SMB CRM Shootout (Test)
Segment: Popular CRM platforms for Small to Medium Businesses

- Salesforce Sales Cloud
- HubSpot CRM Suite
- Microsoft Dynamics 365 Sales

The provided review analyses paint a complex picture of the CRM landscape, with Salesforce Sales Cloud, HubSpot CRM Suite, and Microsoft Dynamics 365 Sales each exhibiting strengths and weaknesses that cater to different business needs. There isn't a clear 'winner', as the ideal choice depends heavily on factors like company size, existing tech stack, budget, and desired feature depth.

Ease of Use & User Experience (UX):

HubSpot consistently receives praise for its intuitive user interface and ease of use, making it a popular choice for beginners and smaller businesses. Reviewers highlight its clean design and straightforward navigation. The free CRM option is also mentioned as a good way to get started and familiarize oneself with the platform. In contrast, Salesforce is often criticized for its dated and clunky UI, with some reviewers describing it as feeling like it's from 2008. While powerful, its complexity can lead to a steep learning curve. Microsoft Dynamics 365 Sales falls somewhere in between. Some reviewers find it less user-friendly for beginners compared to Salesforce, while others appreciate its left-to-right layout and minimal scrolling. The consensus is that Dynamics 365 requires more staff training and customization to achieve optimal usability.

Core Feature Set Comparison:

- **Contact Management:** All three platforms offer robust contact management capabilities. HubSpot is praised for its user-friendly contact storage and the ability to create tasks, log calls, and send emails directly from a contact's record. Salesforce and Dynamics 365 also offer comprehensive contact management, but some reviewers find the data entry process in Dynamics 365 potentially tedious.
- **Deal/Pipeline Management:** HubSpot's sales pipeline management is highlighted for providing a fully functional overview of sales processes and customizable dashboards. Salesforce and Dynamics 365 also offer deal management tools, but Dynamics 365 is noted for having a more streamlined process for converting leads to opportunities.
- **Email Marketing/Automation:** HubSpot excels in this area, offering a wide range of lead capture methods, robust in-platform email marketing tools, and customizable email templates. Salesforce and Dynamics 365 also provide email integration, but HubSpot's email feature is considered more intuitive and advanced than those of its competitors. However, automation features are limited on HubSpot's lower-tier plans.
- **Reporting & Analytics:** All three platforms offer reporting and analytics capabilities. HubSpot's reporting feature is praised for being intuitive to use and offering a wide range of pre-built reports. Salesforce also provides robust reporting and advanced BI tools. Dynamics 365 is noted for its more robust and extensive reporting capabilities, especially when integrated with Power BI.
- **Customization:** Salesforce is renowned for its extensive customization options, allowing businesses to tailor the software to their specific needs and workflows. However, this customization often requires more technical expertise or resources. Dynamics 365 also offers customization capabilities, but it is considered more flexible due to the use of common coding languages like Javascript and .Net. HubSpot offers customization, but advanced customization may require a premium plan.

Pricing & Value Perception:

HubSpot offers a generous free plan, making it an attractive option for small businesses and startups. However, its CRM suite pricing is considered expensive, and hidden costs or upsells are mentioned. Salesforce is also perceived as expensive, especially with add-ons and per-user pricing. Reviewers note that it can be cost-prohibitive for new businesses and that apps that function well with Salesforce come with their own price tag. Dynamics 365 is generally considered more affordable than Salesforce, especially for medium to large businesses. Its pricing is also seen as more straightforward, with no hidden surprises.

Customer Support & Resources:

HubSpot is praised for its 24/7 customer support and quick and responsive live chat. Salesforce has received criticism for its customer support, with some reviewers describing it as not being amazing. Dynamics 365's customer support is not explicitly mentioned in the reviews.

Unique Selling Propositions (USPs) & Key Differentiators:

- **Salesforce:** Extensive customization options, a vast app marketplace (AppExchange), and a comprehensive CRM platform.
- **HubSpot:** Ease of use, a generous free plan, and an all-in-one platform for CRM, marketing, and sales.
- **Microsoft Dynamics 365 Sales:** Flexibility in deployment (cloud, on-premise, hybrid), seamless integration with Microsoft products, and data ownership advantages.

Notable Drawbacks:

- **Salesforce:** Dated UI, high implementation costs, steep learning curve, and inconsistent customer support.
- **HubSpot:** Limited automation features on lower-tier plans, potential email deliverability issues, and HubSpot branding on the free plan.
- **Microsoft Dynamics 365 Sales:** Less user-friendly for beginners, potentially tedious data entry, and more complicated licensing structure.

B Annotation Modules Used in the Study

The following blank annotation modules were used by human annotators and also served as a structural guide for querying the Gemini AI for the two analyzed videos. These modules were designed to standardize the data collection process across all annotators, ensuring that evaluations were made against a consistent set of criteria.

Each module is structured into three main sections:

- **Section I: Feature-Specific Analysis** – This section prompts the annotator to identify whether specific product features were discussed in the video review and, if so, to assign a sentiment score based on an ordinal scale.

- **Section II: Overall Video Review Assessment** – This section gathers judgments on the overall sentiment towards the product, as well as assessments of the video review's presentation quality and the reviewer's characteristics.

- **Section III: Specific Multimedia Event Indicators** – This section focuses on particular cues or elements within the video that might influence the overall interpretation.

The detailed structure and the specific items within each module are presented on the following pages.

Modulo: Smartphone Review (iPhone 16 Series)

ID/Link Video YouTube: ____

ID Annotatore: ___________________________

I. Sentiment ed Analisi per Caratteristica Specifica

Per ogni caratteristica, indicare se è stata discussa e, in caso affermativo, segnare con una X la colonna del sentiment espresso.

Categorie Sentiment: Molto Positivo (MP), Positivo (P), Neutrale (Nt), Misto (M), Negativo (N), Molto Negativo (MN)

Caratteristica/Aspetto Chiave Smartphone	**Discussa? (Sì/No)**	**MP**	**P**	**Nt**	**M**	**N**	**MN**
Usabilità e Dimensioni							
Ergonomia e Maneggevolezza (Pro Max vs Pro)							
Display							
Qualità Generale (Luminosità, Colori)							
Refresh Rate (Fluidità, 60Hz vs ProMotion)							
Prestazioni							
Velocità Generale e Reattività (Processore)							
Gestione Termica / Riscaldamento (vs modello precedente)							
Fotocamere							
Qualità Foto/Video Complessiva (generale)							
Funzionalità Specifiche (es. Zoom, Pulsante Azione per fotocamera)							
Batteria							
Autonomia (Pro Max)							
Autonomia (Pro, e confronto con Pro Max)							
Autonomia (modello "Liscio" / Plus - attese)							
Funzionalità di Gestione Carica (es. limite 80%)							
Design e Materiali							
Materiali (es. Titanio) e Finiture							
Colori e Estetica Generale							
Software e Funzionalità							
Utilità del Pulsante Azione							
Apple Intelligence (Commenti/Aspettative)							
Valore e Posizionamento							
Rapporto Prestazioni/Esigenze Utente (per Pro Max)							
Rapporto Prestazioni/Esigenze Utente (per Pro)							
Valore/Posizionamento iPhone 16 "Liscio"							

II. Valutazione Complessiva della Recensione Video

1. **Sentiment Generale nei confronti dello Smartphone (principale modello recensito):** (Basandosi sull'intera recensione, qual è il sentiment dominante generale dell'autore nei confronti del modello principale trattato?)

   - Molto Positivo □
   - Positivo □
   - Neutrale □
   - Misto □
   - Negativo □
   - Molto Negativo □

2. **Qualità Audio della Recensione (Voce autore, suoni ambiente):**

   - Eccellente □
   - Buona □
   - Media □
   - Discreta □
   - Scarsa □

3. **Efficacia della Presentazione Visiva dello Smartphone nel Video:**
   (Quanto bene viene mostrato lo smartphone? Le dimostrazioni sono chiare visivamente?)

   - Eccellente □
   - Buona □
   - Media □
   - Discreta □
   - Scarsa □

4. **Tono di Voce dell'Autore (Spunto Audio - Dominante):**

   - Entusiasta/Eccitato □
   - Informativo/Oggettivo □
   - Casual/Conversazionale □
   - Critico/Deluso □
   - Monotono □
   - Altro (Specificare): ____________________

5. **Espressività Visiva dell'Autore (Spunto Visivo - Dominante, se applicabile):**

   - Visibilmente Positivo/Coinvolto □
   - Neutrale □
   - Visibilmente Negativo/Deluso □
   - Non Chiaramente Visibile/Applicabile □

III. Indicatori di Eventi Multimediali Specifici

1. **Supporto Visivo/Dimostrativo alle Affermazioni:**
   (Le immagini/azioni mostrate nel video supportano efficacemente le affermazioni verbali dell'autore sulle funzionalità o prestazioni dello smartphone?)

   - Sempre / Quasi Sempre □

- Spesso □
- A volte □
- Raramente □
- Mai / Quasi Mai □
- Difficile da Giudicare / NA □

Modulo: Electric Guitar Review (PRS Silver Sky)

ID/Link Video YouTube: ____

ID Annotatore: _______________________

I. Sentiment ed Analisi per Caratteristica Specifica

Per ogni caratteristica, indicare se è stata discussa e, in caso affermativo, segnare con una X la colonna del sentiment espresso.
Categorie Sentiment: Molto Positivo (MP), Positivo (P), Neutrale (Nt), Misto (M), Negativo (N), Molto Negativo (MN)

Caratteristica/Aspetto Chiave	**Discussa? (Sì/No)**	**MP**	**P**	**Nt**	**M**	**N**	**MN**
Impressione Generale e Confronti							
Impressione Iniziale e Confronto (vs. Stratocaster)							
Manico e Tastiera							
Comfort/Profilo del Manico							
Radius Tastiera e Performance Bend							
Pickups - Caratteristiche Generali							
Carattere Generale (pienezza, precisione)							
Performance con Distorsione							
Hardware							
Precisione del Ponte							
Stabilità Accordatura / Meccaniche Autobloccanti							
Sustain							
Sustain naturale della chitarra							
Suono - Versatilità e Amp Pairing							
Performance con High-Gain							
Sonorità Fender Tweed ’55							
Sonorità Stile Fuchs/Dumble							
Versatilità di Genere Generale (Blues, Rock, Hard Rock)							
Pickups - Qualità Tonali Specifiche							
Tono Pickup al Ponte							
Tono Pickup Centrale							
Tono Pickup al Manico							
"Corposità" / Carattere Simile ai P90							

Caratteristica/Aspetto Chiave	Discussa? (Sì/No)	MP	P	Nt	M	N	MN
Caratteristiche Sonore Complessive							
Timbro Sonoro Complessivo (range, carattere, unicità)							
Feeling e Suonabilità							
Feeling Generale ed Esperienza di Suono							

II. Valutazione Complessiva della Recensione Video

1. **Sentiment Generale nei confronti della Chitarra:**
(Basandosi sull'intera recensione, qual è il sentiment dominante generale dell'autore della recensione nei confronti della chitarra elettrica?)

- Molto Positivo □

- Positivo □

- Neutrale □

- Misto □

- Negativo □

- Molto Negativo □

2. **Chiarezza delle Dimostrazioni Sonore nel Video:**
(Con quale qualità il video cattura il suono della chitarra?)

- Eccellente □

- Buona □

- Media □

- Discreta □

- Scarsa □

3. **Efficacia della Presentazione Visiva della Chitarra nel Video:**
(Quanto bene viene mostrata la chitarra? Gli esempi di suono sono chiari visivamente?)

- Eccellente □

- Buona □

- Media □

- Discreta □

- Scarsa □

4. **Tono di Voce dell'Autore (Spunto Audio - Dominante):**

- Entusiasta / Eccitato □
- Informativo / Oggettivo □
- Casual / Conversazionale □
- Critico / Deluso □
- Monotono □
- Altro (Specificare): ____________________

5. **Espressività Visiva dell'Autore (Spunto Visivo - Dominante, se applicabile):**

- Visibilmente Positivo / Coinvolto □
- Neutrale □
- Visibilmente Negativo / Deluso □
- Non Chiaramente Visibile / Applicabile □

III. Indicatori di Eventi Multimediali Specifici

1. **"[Musica]" (Periodi di Dimostrazione Sonora):**
   Per ogni principale demo sonora, il suono dimostrato supporta efficacemente le affermazioni verbali fatte dall'autore riguardo a quel tono/setting?

- Sempre / Quasi Sempre □
- Spesso □
- A volte □
- Raramente □
- Mai/Quasi Mai □
- Difficile da Giudicare / NA □