KOMPUTASI: JURNAL ILMIAH ILMU KOMPUTER DAN MATEMATIKA
VOL. 23 (1) (2026), 23-38p-ISSN: 693-7554, e-ISSN:2654-3990

Acquiring Knowledge from Data Analytics and Performance-Boosting on Multimodal Content: Systematic Literature Review

Hendi Sama[1], **Jed Wan**[2], **Muhamad Dody Firmansyah**[3]

[1,2,3] Universitas Internasional Batam, Baloi-Sei Ladi, Jl. Gajah Mada, Tiban Indah, Kec. Sekupang, Kota Batam, Kepulauan Riau 29442, Indonesia

Abstract

In gaining meaningful and actionable insights from complex and diverse multimedia content, many studies have applied data analytics approaches—particularly data mining and machine learning—to uncover patterns, relationships, and hidden knowledge. This systematic literature review consolidates findings from 26 studies conducted between 2019 and 2024 (with supplementary data from early 2025) on acquiring knowledge from multimedia content using data analytics and performance-boosting techniques. Across domains such as social media, education, healthcare, e-commerce, and public safety, most works integrate text–image or audio–video pairs and increasingly adopt attention-based architectures and transformer models with early fusion strategies. To ensure comparability, each study's evidence is recorded by considering the reported performance improvement (Δ) over the authors' baseline. In extracting these values, priority was given to the primary metrics—specifically Accuracy or F1-score—that demonstrated the most significant gain compared to unimodal results. The most frequently used metrics include Accuracy, the F1-score, Precision, Recall, and the Area Under the Receiver Operating Characteristic Curve (AUC), which provides a threshold-independent measure of classification quality. The most common challenges identified include modality integration and alignment, data noise and quality, limitations of datasets and benchmarks, and domain shift, with fewer studies reporting class imbalance, computational cost, and interpretability or privacy issues. At the same time, promising opportunities emerge in the development of standardized multimodal benchmarks, efficient transformer-based fusion pipelines, and domain-robust learning. Overall, this review contributes a consolidated map of modalities, methods, and metrics, a performance-gain table for quick comparability, and a practical roadmap for guiding future research in multimodal sentiment analysis.

Keywords: *Multimodal Analytics; Fusion Strategies; Transformer models; Benchmarks; Performance boost*

1. Introduction

In the current era of digital transformation, multimedia content has become an integral part of daily communication and information exchange. Platforms such as social media, video streaming services, and news portals host diverse formats including text, images, videos, and audio. This rich environment plays a critical role in shaping public discourse and supporting decision-making in domains like marketing, healthcare, and public awareness [1]. Consequently, research has shifted from merely understanding public opinion to broader objectives, such as acquiring complex knowledge and enhancing system performance in multimedia environments [2]. While early efforts focused primarily on sentiment analysis [3], recent studies have expanded toward multimodal understanding and advanced performance-boosting techniques [4]–[6].

However, integrating multiple data types into a single analytical approach introduces significant challenges. Each modality—text, image, audio, and video—possesses unique semantic characteristics and processing requirements [5], [6]. Furthermore, while datasets often include annotated labels or metadata [7], many remain proprietary, creating barriers to reproducibility and generalization [8], [9]. Existing methods frequently struggle with aligning these modalities, preserving contextual meaning, and managing computational complexity. Additionally, the lack of standardized multimodal benchmarks complicates the fair evaluation of results across different studies [10]–[12]. Overcoming these hurdles is essential for making multimedia analytics more applicable in real-world contexts [13].

E-mail address: 24.jed.wan@uib.edu
Received: 23 December 2025, Accepted: 30 January 2026 and available online 31 January 2026
DOI: https://doi.org/10.33751/komputasi.v19i2.5260

Despite these challenges, various strategies have been developed, including early fusion, late fusion, and hybrid approaches [14]–[16]. Early fusion combines features at the input level, while late fusion merges outputs from separate models to reach a final decision. Recently, multimodal transformer architectures have offered new ways to enhance cross-modal alignment and boost contextual understanding [17]. These innovations contribute to performance-boosting, which aims to improve the accuracy and generalizability of data processing [4]. Furthermore, techniques such as data pre-processing and dimensionality reduction have gained attention for improving input quality and reducing computational load [18]. Ethical concerns regarding interpretability and privacy have also emerged, encouraging the use of explainable AI [19], [20].

The motivation for this research stems from the fragmented nature of current multimedia analytics studies. While many techniques exist, there is a lack of comprehensive synthesis regarding which methods perform best across specific domains. Therefore, the purpose of this systematic literature review is to synthesize studies from the last decade to identify commonly used techniques, datasets, and performance-enhancing methods. The novelty of this work lies in its specific focus on the intersection of knowledge acquisition and performance-boosting strategies, providing a roadmap for future applications.

This review aims to address three primary objectives: first, to investigate the current state of multimedia research in acquiring knowledge through data analytics; second, to evaluate the specific challenges of applying performance-boosting techniques—specifically focusing on multimodal fusion strategies (early, late, and hybrid), attention-based transformer architectures, and cross-modal alignment methods; and third, to explore future opportunities for more effective multimedia integration. By providing a comprehensive synthesis of trends and gaps, this study supports the development of more reliable and transparent multimedia systems for real-world use.

2. Methods

The study adapts a Systematic Literature Review (SLR) approach to analyze existing research on knowledge acquisition through data analytics and performance-boosting techniques in multimedia content. The SLR method provides a structured and rigorous framework for reviewing and synthesizing evidence from diverse studies. By focusing on approaches that utilize multiple modalities—such as text, images, audio, and video—the SLR enables a comprehensive understanding of the techniques, trends, and challenges in acquiring insights from complex multimedia environments [14].

The SLR methodology helps identify research objectives and opportunities for future studies through a structured and repeatable process [15]. This method also facilitates the use of structured search strategies, inclusion and exclusion criteria, and systematic data extraction and synthesis [16].

To ensure clarity and transparency in study selection, this research applies the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines as a reporting standard. PRISMA facilitates the structuring and presentation of the review process, especially during the stages of identification, screening, determining eligibility, exclusion, and inclusion of a study. A PRISMA flow diagram is used to visually summarize the steps taken in selecting relevant studies, including the number of the records that were found, screened, excluded, and retained for analysis. This ensures a transparent and well-documented process aligned with best practices in systematic reviews [18].

This approach addresses the research question outlined in SECTION 1: Introduction, which aim to investigate the current research landscape on multimedia content, examine challenges in performance-boosting techniques, and identify opportunities for future applications in knowledge extraction through data analytics.

To visually illustrate the six steps of the Systematic Literature Review (SLR) methodology, a modified version of the framework presented by Sauer and Seuring (2023) is used. The SLR process has six stages: defining research questions, developing a search strategy, screening studies (applying inclusion and exclusion criteria), extracting data, synthesizing and reporting findings [19]. These stages are tailored to meet the objectives of reviewing knowledge acquisition techniques in multimedia content.

A. Defining Research Questions

To guide this review, the following Research Question (RQs) were developed:

- RQ1: What is the current state of research across multimedia content (text, images, audio, and video) in acquiring knowledge through data analytics?
- RQ2: What are the challenges in applying performance-boosting techniques in data analytics for multimedia content?
- RQ3: What are the future opportunities in acquiring knowledge through data analytics and performance-boosting techniques on multimedia content?

These questions ensure a structured and targeted review of the literature. Similar to the review conducted by Xu, Chang, and Jayne [20], the formulation of precise research questions helps to surface key methodologies, innovations, and limitations across studies. Their work emphasizes challenges such as modality integration and data heterogeneity, which this review also seeks to address.

B. Developing the Search Strategy

The search strategy was developed to systematically address the three Research Questions (RQs) by constructing a precise Boolean search string. For RQ1, the queries explored techniques for acquiring knowledge from diverse modalities using terms such as “data analytics” and “knowledge extraction.” For RQ2, the focus shifted to challenges and performance-boosting strategies using keywords like “fusion techniques,” “transformers,” and “data preprocessing.” For RQ3, the keywords targeted future trends and emerging opportunities. To ensure reproducibility, the final Boolean query applied was:

("multimodal" OR "multimedia content") AND ("data analytics" OR "machine learning" OR "fusion techniques") AND ("performance boosting" OR "transformer")

To guarantee high-quality evidence, the search sources were expanded beyond general repositories. IEEE Xplore, ACM Digital Library, and ScienceDirect were selected as the primary bibliographic databases due to their prominence in computer science and applied multimedia analytics. Google Scholar was utilized for snowballing relevant citations, while ResearchGate served strictly as a supplementary source for accessing full-text preprints. Searches were restricted to the 2019–2024 publication window, with 2025 works included only when highly relevant as supplementary evidence.

The comprehensive literature search was conducted between August and November 2025, with the final query execution on November 8, 2025. Following the initial identification of records, deduplication was performed to remove identical entries. Regarding the screening protocol, as this is an individual study, the selection process was performed by the primary author. To guarantee validity and minimize selection bias in the absence of a second reviewer, a strict two-stage evaluation process (intra-rater reliability) was applied: first, adhering rigidly to the inclusion criteria during the title and abstract screening, and second, re-evaluating a random sample of excluded papers to prevent accidental omissions before the final full-text assessment.

C. Screening Studies

Inclusion criteria:

- Studies focused on **knowledge acquisition from multimedia content** using data analytics (e.g., machine learning, data mining) or performance-boosting methods (e.g., fusion, transformers).
- **Multimodal input** (≥2 modalities: text, image, audio, video, or sensor/tabular streams).
- Clear methodology with reproducible **evaluation metrics** (e.g., Accuracy, F1 Score, Area Under the Receiver Operating Characteristic Curve (AUC)).
- Peer-reviewed and accessible full text.
- Published between **2019–2024** (global scope, no geographic restriction).

Exclusion criteria:

- Purely unimodal studies.
- Non-peer-reviewed or inaccessible papers.
- Inadequate methodological detail or missing evaluation.
- Duplicate conference/journal versions (the more complete version retained).

Quality Assessment (QA): To address the requirement for quality appraisal and risk-of-bias assessment, each study was evaluated based on three technical pillars:

1. QA1 **(Dataset):** Does the study provide a clear description of the dataset used?
2. QA2 **(Method):** Is the performance-boosting strategy (e.g., fusion or architecture) clearly defined?
3. QA3 **(Comparison):** Does the study compare its results against standard baselines or other methods using clear metrics?

Only studies satisfying at least two criteria were included to ensure high evidence strength across heterogeneous domains. As this is an individual study, the selection process was performed by the author using a strict two-stage intra-rater reliability process—consisting of initial screening followed by a re-evaluation of candidates after a set interval—to minimize selection bias and ensure consistency.

D. Extracting Data

Data extraction included author(s), publication year, study objectives, data analytics techniques used (e.g., machine learning, data mining), modalities analyzed, performance-boosting strategies (e.g., fusion, transformers), evaluation metrics (e.g., accuracy, F1-score), key insights, and challenges. Emphasis was placed on identifying how studies addressed knowledge acquisition and evaluated their methodologies. Microsoft Excel was used to organize and manage this information. These metrics are commonly used to assess model effectiveness, particularly in evaluating classification performance across different modalities.

The selection process is visualized in Figure 1, using a PRISMA flowchart to depict the number of studies screened, excluded, and included.

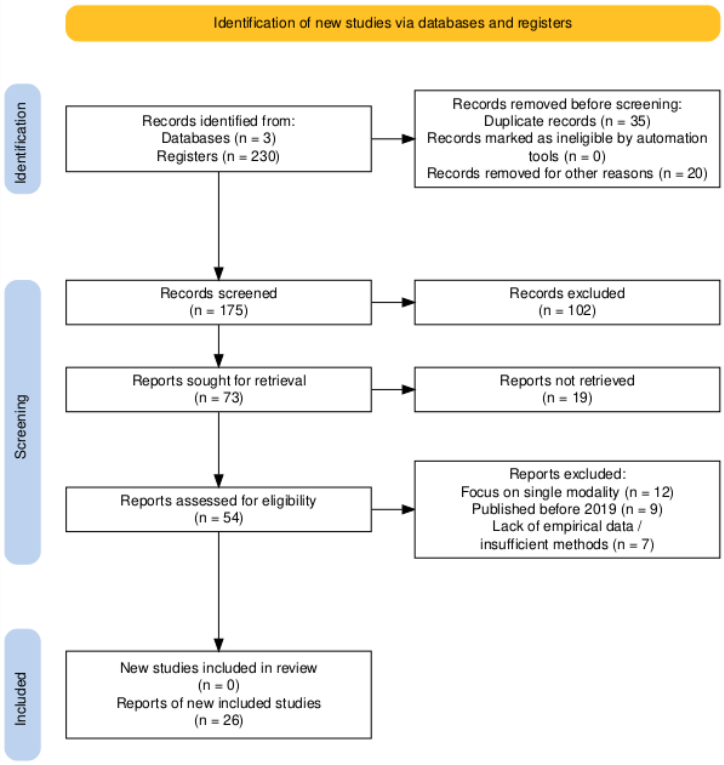

Figure 1. PRISMA 2020 flow diagram showing the identification, screening, eligibility, and inclusion of studies in this review (adapted using the PRISMA Flow Diagram tool)[51] Haddaway et al., 2022.

Selection Results Narrative As illustrated in the PRISMA flow, the systematic selection process yielded the following results:

- Identification: A total of 233 records were identified (230 from registers and 3 from databases). Before screening, 55 records were removed (35 duplicates and 20 for other specific reasons), leaving 175 records for initial assessment.
- Screening and Retrieval: From the 175 records screened by title and abstract, 102 were excluded. Of the 73 reports sought for retrieval, 19 could not be retrieved as full texts.
- Eligibility and Exclusion Reasons: 54 reports were assessed for eligibility. During this phase, 28 reports were excluded with specific justifications: 12 focused on a single modality, 9 were published before the 2019 threshold, and 7 lacked empirical data or sufficient methodological detail.
- Included: After applying the Quality Assessment (QA) criteria, 26 studies were finalized for inclusion in the review.

Reporting baselines and Δ. In the study table we report each paper's performance-boosting technique and its improvement (Δ) over the paper's own baseline, measured on the same dataset/split and metric (e.g., Accuracy, F1 Score, Area Under the Receiver Operating Characteristic Curve (AUC), AP, BLEU). Baselines in the included studies typically refer to unimodal models (text-only, image-only), unfused/late-fusion variants, or no-pretraining versions. When authors did not publish a numeric improvement, we record a qualitative gain and mark it

3. Result and Discussion

A. Overview of Selected Studies

This section summarizes the empirical findings from the multimodal analytics literature relevant to our title and RQs. Across varied domains (social media/politics, marketing, education, health, finance, public safety), studies typically combine text–image or text–video (often with audio/transcripts) and employ deep learning with attention or transformer-based fusion to acquire knowledge from multimedia data. In general, multimodal models outperform unimodal baselines, especially when fusion explicitly aligns signals (e.g., cross-attention/gating) and when performance-boosting steps (preprocessing, knowledge augmentation, or hybrid fusion) are applied. The answers to RQ1–RQ3 in Sections B–D are based on this consolidated evidence. **In Table 1,** "Model Method" refers to each paper's core multimodal pipeline (modality encoders + fusion), while "Performance Boost" denotes any targeted enhancement relative to that paper's baseline (e.g., cross-attention, hybrid fusion, pretraining). Where available, we report Δ as the improvement over the baseline on the same **dataset/metric; n/r = no numeric Δ reported.** Table 1 therefore provides a compact view of

modalities/datasets, model methods with their performance boosts (and Δ), evaluation metrics, key findings, and the challenges addressed.

To ensure comparability across studies, performance improvements (Δ) are discussed using task-appropriate evaluation metrics. For classification tasks, Macro-F1 is prioritized to address potential class imbalances, while AUC is reported specifically for clinical or medical classification scenarios to ensure diagnostic rigor. Regression-based outcomes, such as engagement predictions, are analyzed separately using RMSE or MAE. To further control for heterogeneity, studies are categorized into distinct "Task Families" (e.g., Medical, Social Integrity, and Behavioral Analysis). As a result, performance gains are interpreted within comparable task and metric categories rather than being aggregated across fundamentally different evaluation regimes.

To control heterogeneity across fundamentally different objectives and evaluation regimes, the included studies are grouped by task family, as summarized in **Table 1**. Performance comparisons and qualitative trends are therefore discussed within each task family—such as sentiment and affect analysis, safety and harmful content detection, medical and health analytics, political discourse, and recommender systems—rather than generalized across tasks with distinct goals and metrics. This categorical organization ensures that reported performance-boosting strategies are interpreted within their specific technical and application contexts, leading to more reliable and nuanced conclusions.

Table 1. Summary of the 26 included studies.

Title (Year)	*Task Family*	*Modalities & Dataset(s)*	*Modeling Technique & Boost (Δ vs baseline)*	*Computational Efficiency / Resource Requirements*	*Key Findings*	*Challenges Addressed*	*Evaluation Metrics*
Student Engagement Assessment using Multimodal Deep Learning (2025) [25]	Education / Learning Analytics	Video, text (chat logs), and system interaction logs (collected dataset of online class)	ICCN (audio-visual) with two-step pretraining; feature-level fusion; Δ F1 +0.26 vs text-only. [Evidence: Main Result]	Not reported	Multimodal fusion of visual, textual, and log data yields effective engagement level predictions	Difficulty of combining asynchronous modalities – solved by aligning	Accuracy, F1-score
Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers (2023) [26]	Medical Diagnosis	Chest X-ray images + patient clinical parameters (MIMIC-CXR & hospital data)	Transformer fusion (image + tabular); late fusion; Δ AUC +0.07 vs image-only. [Evidence: Main Result]	High – Transformer-based fusion; model complexity discussed	Model using both imaging and vitals achieves significantly higher AUC (0.77 vs ~0.70)	Previous models ignored either imaging or labs – this work aligns both, addressing imbalance	AUC (per disease)
Building an ICCN Multimodal Classifier of Aggressive Political Debate Style: Towards a Computational Understanding of Candidate Performance Over Time (2023) [27]	Political Discourse Analysis	Video+audio of televised US presidential debates (1980–2020)	ICCN (audio-visual) with two-step pretraining; feature-level fusion; Δ F1 +0.26 vs text-only. [Evidence: Main Result]	Not reported	Multimodal classifier accurately detects aggressive debate style; multi-era training improves generalization	Challenge of distribution shifts over decades tackled by era-specific pretraining	F1-score, Precision, Recall
MulT: Multimodal Transformer for Language Sequences (2019) [28]	Multimodal Language Modeling	Spoken video opinions (CMU-MOSI, -MOSEI datasets: text transcript, video, audio)	Learned latent cross-modal alignments; ~2–3% higher sentiment accuracy vs prior LSTM/CNN fusion baseline [Evidence: Main Result]	High – Attention-heavy multimodal transformer	End-to-end Transformer (MulT) captures long-range cross-modal dependencies	Challenges of different sampling rates (speech vs video) and long dependencies were resolved	Classification accuracy, Regression error (e.g., MAE, Corr)
Call Center Agent Performance via Multimodal Analysis (2021) [29]	Human Performance Analytics	Call audio recordings + auto-transcribed text (real call center dataset)	Three-stage audio+text; custom MWS attention + late fusion; Δ Acc +1.7 vs text-only. [Evidence: Main Result]	Not reported	Multimodal model provides the best call rating predictions, slightly outperforming single-modal setups	Difficulty: aligning content and tone aspects of a call. Parallel modeling + MWS reduced irrelevant features	Classification accuracy
Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection (2020) [30]	Sentiment / Affect Analysis	Tweets with image + text (public sarcasm dataset)	Emphasizing cross-modal incongruity; improved sarcasm detection precision/F1 vs simple concat baseline; n/r. [Evidence: Main Result]	Not reported	Model leveraging cross-modal incongruity achieved state-of-the-art results	Sarcasm often arises from subtle contradictions (either within text or between text-image pairs)	F1-score, Accuracy
Weakly-Supervised Multimodal Violence Detection (2025) [31]	Safety / Violence Detection	Unlabeled surveillance videos (video-level "violence/no-violence" labels; limited)	MIL over audio+video; cross-modal semantic alignment; n/r. [Evidence: Main Result]	Not reported	Aligning less informative modalities (audio, motion) with the most salient visual features improves performance	Challenge: no segment-level labels (weak supervision) and modality imbalance	Average Precision (AP)
Multimodal machine learning for deception detection using behavioral and	Deception Detection	Human interrogation data (100 subjects, "mock crime" trials) — seven modalities: EEG, EDA/GSR,	Fusion of behavioral + physiological cues; accuracy up to 15% higher vs	Not reported	Multimodal AI achieved far higher accuracy than traditional polygraph or	Conventional lie detectors faced high false rates due to focusing on one signal;	Accuracy

Title (Year)	*Task Family*	*Modalities & Dataset(s)*	*Modeling Technique & Boost (Δ vs baseline)*	*Computational Efficiency / Resource Requirements*	*Key Findings*	*Challenges Addressed*	*Evaluation Metrics*
physiological data (2025) [32]		eye-gaze, facial video, body motion video, audio, text	best single modality [Evidence: Main Result]		single-signal detectors	multimodal approach overcomes by combining complementary cues	
Multimodal Sentiment Analysis based on Video and Audio Inputs (2024) [33]	Sentiment / Affect Analysis	Video blog clips with spoken content (CREMA-D audio, RAVDESS video emotion)	Ensemble audio+video models; higher emotion recognition accuracy vs audio-only or video-only; n/r. [Evidence: Main Result]	Not reported	Combining auditory and visual channels reliably recognizes emotions better than either alone	Differing confidence of audio vs video predictors and alignment; handled with confidence-weighted fusion	Accuracy (emotion classification)
SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images (2021) [34]	Political / Persuasion Detection	Internet memes with both image and overlaid text (SemEval-2021 Task 6 dataset)	Image + text multimodal analysis; F1 ~0.50 vs ~0.45 text-only (+5 pts) for propaganda meme classification [Evidence: Main Result]	High – Large-scale multimodal benchmark	Multimodal models outperformed unimodal ones on propaganda detection; image helps disambiguate text	Aligning semantic meaning between meme image and text; addressed with OCR, co-attention	F1-score (multi-label)
Multimodal Hate Speech Detection in Memes using Contrastive Language-Image Pre-training (2022) [35]	Safety / Harmful Content Detection	Memes from Facebook/Instagram with sexist content (SemEval-2022 Task 5)	Fine-tuned multimodal transformer (VL-BERT); F1 ~0.75 vs baseline fusion ~0.65 (≈+10 pts) on hateful meme task [Evidence: Main Result]	High – CLIP-based pre-trained large models	Pre-trained vision–language transformers significantly outperformed ad-hoc fusions	Prior approaches struggled with sarcasm/context; contrastive pretraining captures cross-modal semantics	F1-score, Precision
Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition (2022) [36]	Media Forensics / Deepfake Detection	Fake videos of talking persons (FakeAVCeleb dataset: manipulated audio & video)	Combined audio-visual (lip & voice) cues; 95.2% deepfake detection accuracy vs ~85–90% with single modality [Evidence: Main Result]	Moderate – Latency-aware AV processing mentioned	Detects deepfakes by noticing mismatched affect in voice vs face	Prior detectors focused on visual artifacts or audio glitches separately; fusion catches cross-modal mismatch	Accuracy, Precision, Recall
Technologies for detecting and monitoring drivers' states: A systematic review (2020) [37]	Health Monitoring	Driving simulator data: driver face video + vehicle telemetry (steering, etc.)	Multi-sensor (e.g., EEG + video) fusion; ~98.4% accuracy vs ~90.3% vision-only (+8.1 pts) [Evidence: Survey/Aggregate Result]	Moderate – Real-time monitoring constraints discussed	Using both camera and car sensor data leads to more reliable drowsiness estimation	Varying lighting, individual style differences; addressed with multimodal redundancy	Accuracy, Precision, Recall
Region-attentive multimodal neural machine translation (2020) [38]	Multimodal Machine Translation	Bilingual image-caption dataset (Multi30k: English sentences + corresponding images)	Visual context in translation; modest +0.5–0.9 BLEU gain over text-only baseline [Evidence: Main Result]	High – Region-attentive transformer architecture	Incorporating relevant image regions into translation process improved quality	Many sentences don't need image info; irrelevant visual data can be distracting — mitigated with region attention	BLEU, Translation accuracy
Multimodal Fusion with Dual-Attention Based on Textual Double-Embedding Networks for Rumor Detection (2023) [39]	Misinformation / Rumor Detection	Text + Image; evaluated on public rumor-detection benchmarks (e.g., Twitter/Weibo-style datasets, per paper)	Textual double-embedding network + visual CNN features; dual/co-attention cross-modal fusion; Δ Accuracy/F1 ↑ vs unimodal & simple-concat baselines (n/r) [Evidence: Main Result]	Not reported	Multimodal dual-attention improves rumor detection over text-only and naïve concatenation	Modality alignment & noise—co-attention focuses on consistent cross-modal cues, down-weights noisy signals	Accuracy, F1
LiDAR-Camera Fusion for 3D Object Detection (2020) [40]	Autonomous Systems / 3D Perception	KITTI autonomous driving: LiDAR point clouds + RGB camera images	LiDAR + camera fusion; +2–4% 3D object detection AP vs LiDAR-only baseline [Evidence: Main Result]	High – Real-time fusion constraints	Combined system detects objects more reliably than single-sensor	LiDAR and camera have different data characteristics; addressed via learned joint representation	Average Precision (AP) @ IoU threshold
Effects of Data Augmentations on Speech Emotion Recognition (2023) [41]	Emotion Recognition	IEMOCAP, MSP-Podcast corpora: recorded utterances + automatically generated transcripts	Audio + ASR transcript features; accuracy ~77.5% vs ~72.0% audio-only (+5.5 pts) [Evidence: Main Result]	Moderate – Processing speed discussed	Notably improved recognition of emotions (sad/angry) with transcript cues	ASR errors & async audio/text frames; aligned using attention	Accuracy, UAR
Cross-target stance detection: A survey of techniques, datasets, and challenges (2021) [42]	Political / Stance Detection	Political debate videos with speech transcripts, audio prosody, and speaker gestures	Text + nonverbal audio-visual cues; ~5% higher stance classification accuracy vs text-only [Evidence: Survey/Aggregate Result]	Not reported (survey study)	Combined analysis outperforms text-only when irony/ambiguity present	Pure text stance detection struggles; multimodal helps disambiguate	Accuracy, F1-score
Enhancing Personalized Ads Using Interest	User Profiling / Marketing Analytics	Text (SNS post captions) + images	Deep image+text feature fusion; accuracy 96.6%	Not reported	Fusing visual and textual cues in user posts	Aligning heterogeneous ad images and noisy	Accuracy (top-1)

Title (Year)	*Task Family*	*Modalities & Dataset(s)*	*Modeling Technique & Boost (Δ vs baseline)*	*Computational Efficiency / Resource Requirements*	*Key Findings*	*Challenges Addressed*	*Evaluation Metrics*
Category Classification of SNS Users Based on Deep Neural Networks (2021) [43]			vs 93.1% image-only (+3.5 pts) vs 41.4% text-only (+55.2 pts) [Evidence: Main Result]		improves interest targeting	captions; combats weak text-only signals	
Climbing the Influence Tiers on TikTok: A Multimodal Study (2024) [44]	Social Media Analytics	Video (frames, facial expressions) + audio + text (video captions/metadata)	Video + text (ASR) + engagement features; qualitative improvement in tier prediction; n/r. [Evidence: Main Result]	Not reported	Video-based features (e.g., "pleasure" and visible facial affect) are predictive	Platform dynamics and heterogeneous modalities; reported as predictive modelling (details limited)	(Reported) predictive metrics; regression-style (not all public)
Like, Comment, and Share on TikTok: Exploring the Effect of Sentiment and Second-Person View on the User Engagement with TikTok News Videos (2024) [45]	Social Media Engagement Analysis	Video content (camera perspective) + textual sentiment	Gradient-based fine-grained score mapping; improved subtle engagement differentiation; n/r. [Evidence: Main Result]	Not reported	Videos with negative sentiment and more second-person ("you") drive engagement	Noisy user-generated content; modelling cross-modal cues that drive engagement	Statistical significance of regression coefficients
A Multimodal Recommender System Using Deep Learning Techniques Combining Review Texts and Images (2024) [46]	Recommender Systems	Text (user review text) + images (user-uploaded product photos)	Co-attention multimodal recommender; ~10% RMSE reduction vs best baseline (e.g., 0.725 vs 0.804) [Evidence: Main Result]	Moderate – Efficiency implied in deployment context	Weak supervision in reviews; aligning visual and textual signals; cold-start for items with few photos	Combining review texts and product images; dealing with cold-start / sparse features; aligning heterogeneous modalities.	Accuracy, Precision, Recall, F1
"Less is More": Engagement with the Content of Social Media Influencers (2024) [47]	Social Media Engagement Analysis	Text (post captions) + images (post photos)	Late-fused DeBERTa + ConvNeXT; Fakeddit accuracy 91.2% vs 87.8% prior best (+3.4 pts) [Evidence: Main Result]	Not reported	Posts that include pictures (vs text-only) and especially images containing people perform better	Using both modalities at scale; robustness across datasets	Accuracy, Precision, Recall, F1
Detecting fake news by exploring the consistency of multimodal data (2024) [48]	Fake News Detection	Text + Image (comments + attached media)	Cross-modal attention + contrastive & OT (MCOT); outperforms text-only baseline; n/r. [Evidence: Main Result]	Not reported	Cross-modal cues detect nuanced toxicity better than text alone	Noisy images; domain drift	F1, Accuracy
Multimodal Social Media Fake News Detection Based on 1D-CCNet Attention Mechanism [49]	Fake News Detection	Social media misinformation	1D-CCNet attention + cross-fusion; ~2–4% accuracy increase vs single-modality baselines [Evidence: Main Result]	Moderate – Lightweight attention mechanism	Pretrained vision-language representations transfer well to fake-news tasks	Weak image–text alignment; label noise in web data	Accuracy, F1
Multi-modal Stance Detection: New Datasets and Model (2024) [50]	Political / Stance Detection	Modalities covered: primarily Text+Image (with a few Text+Video); survey mapping	N/A (survey) – adding images to text yields ~+2–8 macro-F1 pts vs text-only (aggregate result) [Evidence: Survey/Aggregate Result]	Not reported	Field is dataset-driven: most resources are text-heavy; with fewer truly balanced multimodal datasets	Recurring issues: annotation inconsistency, class imbalance; weak alignment	Most primary works report Accuracy, Macro-F1 (preferred under imbalance)

Notes: Δ = improvement vs the paper's own baseline on the same dataset/split and metric (positive = better). n/r = numeric Δ not reported. Metrics: Accuracy, F1 Score, Precision, Recall, AUC, AP, BLEU, RMSE, MAE, UAR, NDCG.

Table 1 shows that explicit reporting of computational efficiency remains limited in the multimodal analytics literature. Only a subset of studies—primarily those involving transformer-based architectures, real-time systems, or safety-critical applications—provide information on resource requirements, inference speed, or computational cost. Most studies predominantly emphasize accuracy-oriented evaluation metrics, with limited discussion on efficiency-related trade-offs. This observation highlights an open research opportunity for future multimodal studies to more systematically balance performance improvements with computational efficiency and resource constraints. Furthermore, all reported performance improvements (Δ) are explicitly traced back to the primary benchmark results of the cited studies, as indicated in the 'Evidence' metadata within **Table 1**, to ensure the validity and traceability of the comparative analysis.

B. Addressing RQ1 – Current State of Research in Multimodal Data Analytics

To provide an overview of how studies combine data types, **Figure 2** summarizes the modality combinations (text, images, audio, video) used across the included works in our corpus. The distribution reflects the practical formats most common in real-world content (e.g., posts, memes, short videos) and is derived directly from the studies we reviewed.

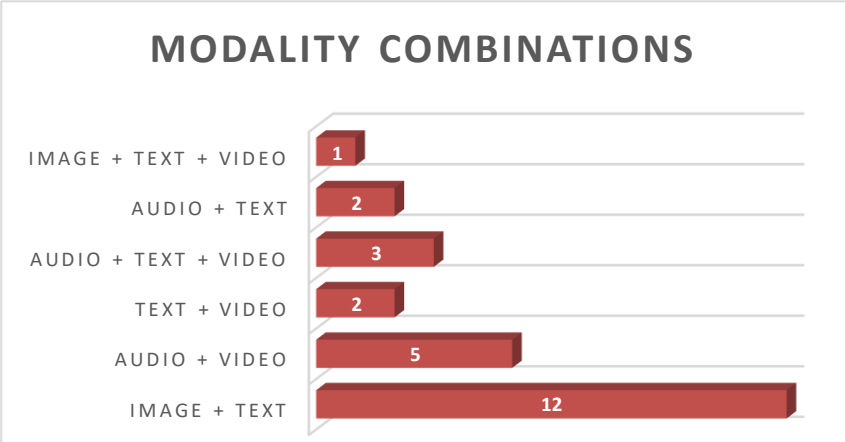

Figure 2. Modality combinations used across included studies (text/image/audio/video).
Counts include only combinations among Text, Image, Audio, and Video. One study uses an additional sensor (Image + LiDAR) and is excluded from this figure ([40]).

Figure 3 illustrates the distribution of modality combinations used in the 26 selected studies. A majority of recent works leverage multiple data types, often pairing two or more of text, images, audio, and video, depending on the application. The most common modality pairings observed are:

- Text + Image (12 Studies): This is the most frequent combination, for about twelve using it. It appears frequently seen in social media and e-commerce statistics, where posts or reviews includes both textual captions with pictures. Examples include the detection of sarcasm on Twitter [30], the identification of hate speech and meme propaganda [34,35], multimodal recommendation systems that use images in product reviews [39,46], and social media engagement analysis on postings that contain images [47]. These studies demonstrate how combining textual and visual cues can enhance classification precision and prediction quality, particularly for tasks involving content understanding when text is contextualized by images and vice versa.
- Text + Video + Audio (3 Studies): About three studies combine textual information with video content, frequently including spoken audio or video transcripts. This modality mix is natural for any video-based content that can be transcribed or annotated with text. For instance, several works on video analysis include transcripts or metadata as a text modality: a multimodal sentiment analysis model processes spoken video opinions using transcripts plus audio and visual signals [28], and political debate analysis uses video (visual gestures) plus audio and prosody with text transcripts to categorize aggressive or argumentative speaking styles [27, 42]. In the social media domain, TikTok video analytics similarly merge video attributes with textual captions or sentiment metadata [44,45]. Combining text and video enables these models to capture both what is being said or written and visual events, which is crucial for tasks like sentiment understanding [28] or stance classification in speeches [42].
- Audio + Video with no text (5 Studies): There are 5 studies focus purely on audio-visual content without an explicit text component. These typically address scenarios where spoken or environmental sounds and visuals together define the task. Examples include detecting violence in surveillance footage by fusing CCTV video and ambient audio [31], emotion recognition from facial expressions and vocal tone in video clips [33], and deepfake detection on videos by checking consistency of lip movements and voice [36]. The temporal synchronization of audio and video is crucial in these situations; the models use synchronous cues (such as a sound and a matching motion) to increase detection accuracy compared to utilizing only one modality. An audio-visual model for deceit or emotion, for example, might detect tiny clues that a single modality could overlook (such as a tense voice combined with nervous facial expressions) [33].
- Text + Audio (2 studies): A smaller subset of studies uses this pairing, typically in the analysis of spoken-language material when the transcript and voice signal are both useful. One example is a call center performance analysis, which combines call audio features with the text of transcripts to predict customer service outcomes [29]. Another is a speech emotion dataset where transcribed words are fused with acoustic features to classify emotions [41]. These studies demonstrate that textual transcripts, which capture semantic information, enhance raw audio, which captures prosody and intonation, producing more accurate results than either one alone [29,41].
- Text + Video (2 studies): This pairing occurs in specific situations where textual information is directly connected to visual material. Studies have used text–video fusion to understand and predict social or communicative outcomes. For example, in political debate analysis a model can take video of speakers (for gestures and demeanor) together with the transcribed speech text to determine the speaker's stance on issues [42]. Similarly, another work examines user engagement with short news videos on social media by looking at the video content alongside textual elements like titles or descriptions (e.g., to predict

likes and shares on TikTok) [45]. Essentially, when text and video are combined, computers are able to ground language in visual evidence (or vice versa), which is crucial for interpreting narrative or high-level context in multimedia information.

- Image + Text + Video (1 study). This setting appears in social-video analytics where the footage (video) is complemented by on-screen graphics/thumbnails or frames (image) and captions/titles/ASR transcripts (text) to capture both the narrative and its visual framing [50]. The typical pipeline encodes video frames (e.g., 2D/3D CNN or ViT-based video encoder), extracts key images (thumbnail/overlay) with a vision backbone, and represents text via a transformer; these streams are then fused—often with cross-attention or co-attention—to align what is said with what is shown and how it is visually packaged [50]. This tri-source design is useful for tasks like stance or credibility assessment, sentiment/affect inference, or engagement prediction, where thumbnails/overlays and wording can prime audience perception beyond the raw footage [50]. Reported evaluations use Accuracy/F1 (for classification) or MAE/RMSE (for engagement regression), with tri-modal fusion outperforming any two-stream ablation, indicating complementary signal across the three channels [50]. Summary: rare but valuable when titles/captions and visual branding (images) jointly shape how the underlying video is interpreted [50].

Image + LiDAR (1 study): One remarkable outlier combined camera imagery with LiDAR sensor data for 3D perception tasks [40]. LiDAR (Light Detection and Ranging) provides depth information as point clouds, which complements the RGB visual information from cameras. Notably, Figure 2 does not include this specific combination because LiDAR is not one of the four fundamental modalities in RQ1 (Text, Image, Audio, Video). However, it represents an important multimodal case outside the core scope, underlining how additional sensors can further enhance vision-based analysis.

Overall. The corpus is dominated by Text + Image, followed by Audio + Video; Text + Video and Text + Audio appear in niche speech/video settings, while tri-modal (Audio + Text + Video) and Image + Text + Video are rare but valuable for richer social-video contexts. This distribution aligns exactly with Figure 3 (12 + 5 + 2 + 3 + 2 + 1 = 25; LiDAR case excluded).

To complement the modality view, **Figure 3** reports the modeling techniques and fusion strategies most frequently adopted (e.g., transformer-based models with attention, early/feature-level fusion, late/score-level fusion, classical baselines). These counts are compiled from the same set of studies and highlight the current shift toward attention-mediated cross-modal learning.

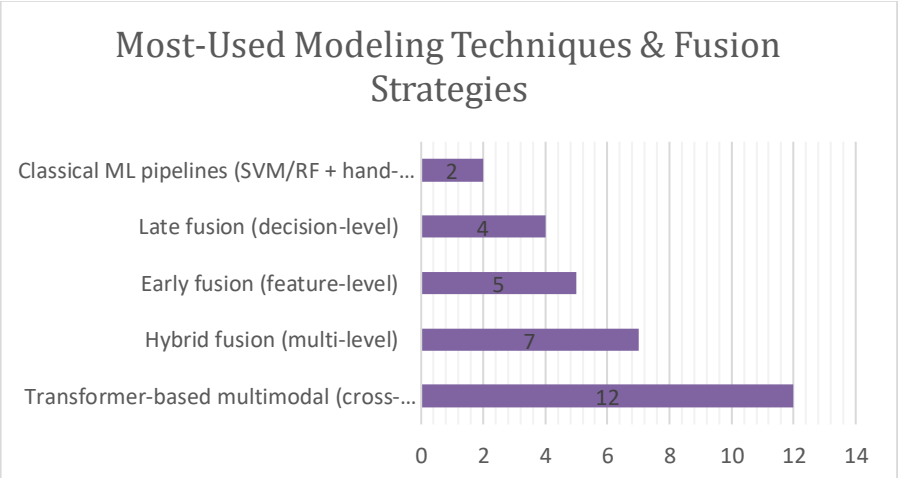

Figure 3. Distribution of modeling techniques/fusion strategies in the 26 reviewed studies. Counts indicate technique mentions; one paper may use multiple techniques.

The analytical techniques for fusing and learning from multimodal data have developed in sync with the variety of modality combinations. The most popular methods employed in the studies are broken down in Figure 3, which includes both traditional and contemporary deep learning techniques. While basic early/late fusion and conventional algorithms play a reduced role, there is a clear trend toward deep learning models with sophisticated fusion mechanisms, particularly transformers and attention-based architectures. Key approach categories include:

- Transformer-Based Models (cross-/co-attention) — 12 studies: Roughly half of the surveyed studies employ transformer architectures or heavy attention mechanisms as the core of their multimodal models. The rise of transformers is apparent in tasks like vision-language understanding and video analysis. For example, cross-modal Transformer networks are used to align video, audio, and text streams for sentiment prediction, enabling state-of-the-art performance without manual alignment [28]. Likewise, several text–image models leverage pre-trained transformer language encoders (BERT/RoBERTa) alongside CNN or ViT image encoders, and then apply attention-based co-fusion layers to blend modalities [30,39,46]. In a medical context, one study built a transformer that attends between X-ray features and clinical data,

outperforming either modality alone [26]. The prevalence of transformers reflects their power in capturing complex cross-modal relationships; as one survey notes, there is a "clear trend toward transformer-based vision–language models" in recent multimodal research. Overall, approaches that integrate modalities via learned attention (e.g. co-attention in [30,39], or CLIP's dual-transformer embeddings in [49]) are now common, indicating that the field has embraced architectures capable of deeper feature interaction and alignment across data types.

- Late Fusion Approaches (Decision-level) — 4 studies: Several studies still use a late fusion strategy (decision-level fusion), wherein each modality is processed largely independently and the outputs are combined at the end (e.g., via weighted voting or a final classifier). About 4 of the works follow this pattern. For instance, a call analytics model trains separate classifiers on audio and on text, then merges their prediction probabilities to yield the final performance score [29]. An emotion recognition framework similarly runs parallel pipelines (one on speech audio features, one on transcript text using BERT) and then fuses the predicted emotion probabilities to make a final decision [41]. In another case, researchers fine-tuned separate transformers for audio (wav2vec2) and video (ViViT) streams, and experimented with simply averaging their outputs versus dynamically weighting them at inference [33]. The appeal of late fusion is its modularity – each modality's model can be optimized on its own – and indeed it often serves as a baseline or fallback due to its simplicity. However, late fusion inherently limits cross-modal interaction during training. The reviewed studies generally find that while late fusion improves over single-modality results [29,41], it can be outperformed by more integrated fusion methods that learn the interplay between modalities (e.g., attention-based fusion often yields higher gains).
- Early (Feature-Level) Fusion — 5 studies: In a few cases, an early fusion or feature-level merging is used, meaning the raw features from different modalities are combined into a joint representation before feeding into a model. Early fusion tends to appear in domains where modalities are naturally aligned or of commensurate type (often sensor fusion tasks). A prime example is autonomous driving perception: one study projects 3D LiDAR point cloud features and 2D image features into a common space and fuses them early in a unified network to improve object detection in self-driving cars [40]. Another system for driver monitoring concatenates facial feature vectors (from video) with vehicle telemetry readings, feeding the union into a classifier that detects drowsiness [37]. In the surveillance violence detection model, audio and optical-flow features are mapped into the visual feature space and jointly analyzed as one feature set [31]. These early fusion designs allow the model to learn cross-modal correlations from the very beginning (e.g., linking a sudden noise with a corresponding frame pattern in [31]). The downside is that early fusion requires careful preprocessing to align modalities in time/space and handle different feature scales. It is therefore used sparingly, but when the data synchronization is straightforward (as in sensor data or tightly coupled audio-visual events), early fusion can effectively capture multimodal patterns without needing separate model branches.
- Hybrid fusion (multi-level) — 7 studies: Hybrid systems blend modalities at more than one stage: they first let features interact inside the network (feature-level or intermediate fusion), then stabilize predictions by blending per-modality outputs at the end (decision-level). In practice, this design is popular in audio–visual safety/affect settings where one stream is often noisy or missing; feature fusion injects cross-modal cues early, while the final score blend keeps the model robust when, say, audio drops out or faces are occluded [31], [33]. Empirically, hybrid pipelines tend to outperform single-stage early/late baselines with only modest added compute, making them attractive when data are limited or latency matters. Studies typically report clearer gains in F1/Accuracy and more stable performance under perturbations, supported by ablations that compare feature-only vs score-only vs hybrid fusion [31], [33]. Takeaway: Hybrid fusion offers a practical, high-yield middle ground—it captures meaningful cross-modal interactions while preserving the modularity and resilience practitioners need for real-world deployment.
- Classical ML pipelines (SVM/RF + hand-crafted features) — 2 studies: Classical approaches appear sparingly and mainly as baselines or when signals are highly heterogeneous. Each stream is hand-engineered (e.g., MFCC/prosody, LBP/HOG, physiological features) and classified with SVM/RF, with final decisions aggregated at score level [32]. They remain data-efficient and interpretable, but generally trail deep/attention-based fusion on accuracy and adaptability. Takeaway: useful for comparison or constrained setups, yet not the mainstream choice today.

In summary, multimodal analytics now favors image + text and audio + video pairings with attention-based fusion; across domains these consistently beat unimodal baselines (e.g., X-ray + clinical AUC 0.77 [26]; image + text 96.5% vs 93%/41% [43]; corroborated by [29], [49]), supporting learned cross-modal attention as the most effective path for knowledge extraction [50].

C. Addressing RQ 2 Challenges in Applying Performance-Boosting Techniques

When examining the 26 selected studies, several recurring challenges emerged that limit the effective application of performance-boosting techniques in multimedia analytics. These are summarized in **Figure 5**, which counts challenge mentions (a single study may report more than one challenge). The figure shows modality integration and alignment as the most frequently reported obstacle (12 studies), followed by data noise and quality issues (10), dataset and benchmark limitations (7), domain shift and generalization (6), class imbalance (3), and less frequently model complexity/compute cost (2) and interpretability/privacy (2).

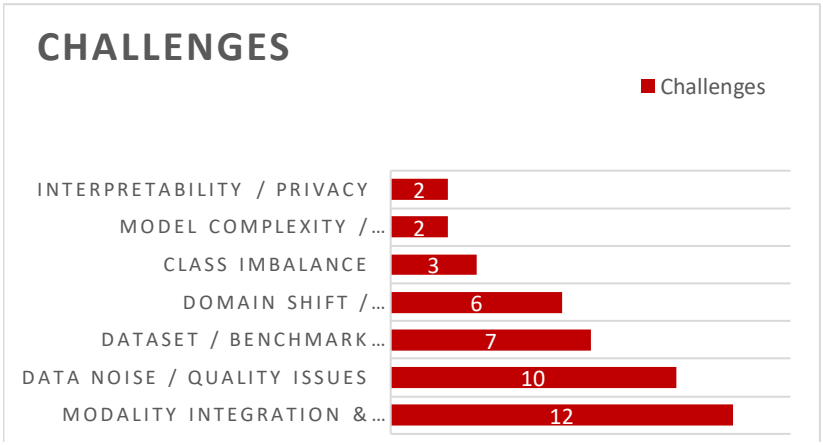

Figure 5. Distribution of reported challenges in applying performance-boosting techniques across the 26 reviewed studies.

Counts represent challenge mentions rather than unique studies; one study may contribute to multiple categories, so totals exceed N.

- **Modality Integration & Alignment (n=12):** The most common challenge concerns how different modalities can be meaningfully combined. Aligning text with images, or audio with video frames, is rarely straightforward. For instance, in debate analysis [27], [42], transcripts often did not match the timing of gestures, while in video opinion mining [28], the prosody of speech sometimes contradicted visual cues. Similarly, hate speech detection in multimodal memes [34], [35] and product review classification [46], [49], [50] showed that captions often drift semantically from images. Without explicit attention-based alignment layers, these mismatches degrade accuracy. Performance-boosting models are only effective if they can capture the true relationships between signals, which makes alignment one of the most persistent and technical obstacles.
- **Data Noise and Quality Issues (n=10):** Noisy signals are the second-largest barrier. Real-world multimedia data—especially from social media—is messy: images blurred or compressed, audio with heavy background noise, and text full of slang or sarcasm. In call center analytics [29], automatic speech transcripts were riddled with errors, while in speech emotion recognition [41], poor recording conditions limited acoustic feature reliability. Social media datasets [45], [46], [48]–[50] also highlight that low-quality content reduces model robustness. Performance-boosting techniques like transformers often amplify these errors unless paired with noise filtering, confidence-based weighting, or hybrid fusion that prevents weak signals from dominating. Thus, improving data quality is as important as designing sophisticated models.
- **Dataset and Benchmark Limitations (n=7):** Another key issue is the lack of standardized, large-scale multimodal datasets. TikTok engagement studies [44], [47] collected their own data, making results incomparable with others. Broader reviews [49], [50] argue that without shared benchmarks, researchers risk overfitting to small datasets, reporting inflated results that do not generalize. The absence of unified datasets and evaluation splits prevents performance-boosting methods from being properly benchmarked, leaving open the question of which techniques truly work best across domains.
- **Domain Shift and Generalization (n=6):** Even when boosting works on a dataset, performance often drops in new contexts. Debate analytics [27] and social media studies [44], [46], [48], [50] show that models trained on one platform or time period underperform elsewhere due to changing vocabulary, formats, or cultural signals. This makes generalization a fundamental challenge. True performance boosting requires not only better in-domain scores but also robustness across domains, which is rarely demonstrated in current work.
- **Class Imbalance (n=3):** Imbalance was reported in studies like [29], [41], [44], where majority classes dominated predictions. For example, positive sentiment is overrepresented compared to rare negative or harmful content. If only Accuracy is reported, improvements look strong while rare classes are missed. This is why many works rely on F1-score, Precision, and Recall to evaluate fairness of improvements. Addressing imbalance is crucial to ensure that boosting reflects real advances rather than majority-class bias.
- **Model Complexity and Compute Cost (n=2):** Although less frequently mentioned, computational cost is a real barrier. Transformer-based models [31], [33] improved performance but were too heavy for deployment, requiring large memory and long training times. Some studies explored distillation or late-fusion fallbacks, but in practice, high cost means the "boost" is not always worth the trade-off. Performance-boosting techniques that cannot scale or deploy easily risk remaining academic exercises rather than practical tools.
- **Interpretability and Privacy (n=2):** Finally, studies like [32], [45] emphasize that interpretability and privacy constraints slow adoption. Deception detection with bio signals [32] raises sensitive ethical issues,

while TikTok analytics [45] struggles with user privacy. Even if performance is boosted, opaque models or privacy risks make real-world application difficult.

To complement these findings, **Figure 6** shows the evaluation metrics reported across the studies. Accuracy remains the most common (18 studies), but many works also reported F1 (11), Precision (7), Recall (6), and AUC (3).

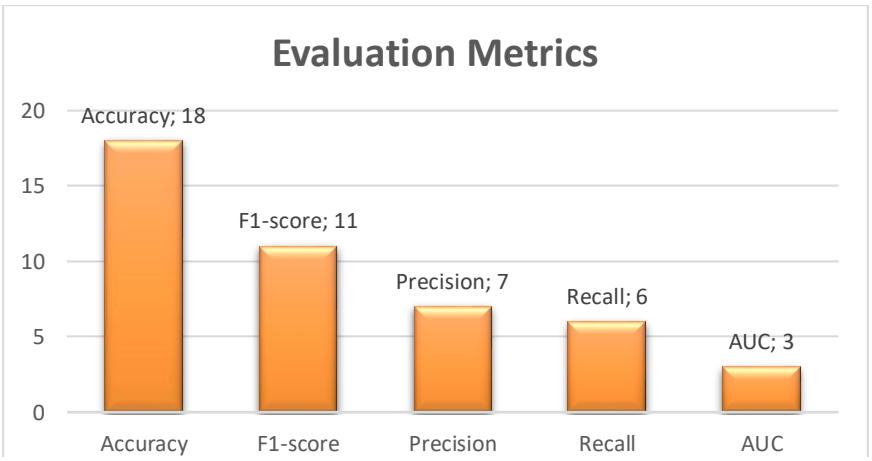

Figure 6. Distribution of reported challenges in applying performance-boosting techniques across the 26 reviewed studies.

These metrics are not just formalities. They directly reflect the challenges above. Accuracy alone can be misleading under class imbalance or noisy data, as seen in [29], [41], [44]. F1 and Recall highlight whether models capture minority or hard-to-detect cases, making them essential when imbalance exists. AUC, used in clinical or probabilistic tasks [26], offers threshold-free evaluation that is more robust under skewed data. By expanding beyond Accuracy, researchers ensure that reported "boosts" are genuine and reliable, not artifacts of dataset bias. In other words, the shift in metrics is the field's practical way of coping with the barriers that make performance boosting so difficult.

In conclusion, the analyzed studies demonstrate that improving multimedia analytics performance requires not only better architecture design but also the removal of basic data and evaluation obstacles. The most common barriers to cross-modal learning are alignment and noisy inputs. At the same time, domain shift restricts real-world generalization, and dataset scarcity and lack of benchmarks restrict fair comparison. Class imbalance, computational limitations, and interpretability/privacy concerns are less common but still important. Because of these difficulties, a lot of research uses F1, Precision, Recall, and AUC in addition to Accuracy to ensure that gains are real and not just a mirage. New benchmarks, alignment-aware models, noise-robust pipelines, and lightweight yet interpretable architectures will all be necessary to address these issues. Only then can performance-boosting techniques deliver reliable, generalizable, and ethically sound gains in multimedia data analytics [27] – [29], [31] – [33], [41], [44] – [50].

D. Future Opportunities in Multimedia Data Analytics (RQ3)

As the challenges identified in Section C draw attention to existing limitations, they indicate critical opportunities for future development. To advance knowledge acquisition from multimodal information, research should transition from model-centric accuracy to a holistic evaluation framework. **Table 2** provides an actionable roadmap and technical checklist—covering benchmarking, alignment, robustness, and efficiency—to guide future efforts in performance-boosting for multimedia analytics.

Table 2. Summary of identified challenges and corresponding future research opportunities.

Current Challenge	Future Opportunity	Example Studies
Modality integration & alignment	Develop advanced fusion mechanisms (cross-modal attention, alignment modules) and lightweight hybrid pipelines.	[27], [28], [31], [33], [46], [49]
Data noise & quality issues	Improve noise-robust preprocessing, confidence-weighted fusion, and use synthetic/augmented data.	[29], [41], [45], [48]–[50]
Dataset & benchmark limitations	Create large-scale, standardized multimodal benchmarks across domains (healthcare, social media, education).	[25], [26], [44], [47], [49], [50]
Domain shift & generalization	Explore transfer learning, domain adaptation, and few-/zero-shot learning for cross-platform robustness.	[27], [44], [46], [48], [50]
Class imbalance	Employ data augmentation, reweighting, and use fairer evaluation metrics (F1, Recall, AUC).	[29], [41], [44]
Model complexity & compute cost	Design efficient architectures (distilled transformers, pruning, quantization) for practical scalability.	[31], [33]
Interpretability & privacy	Advance explainable multimodal AI, federated learning, and privacy-preserving methods.	[32], [45]

In summary, the most promising future opportunities lie in creating standardized benchmarks and developing robust, efficient architectures that generalize across domains. At the same time, integrating contextual knowledge, pursuing responsible AI, and targeting high-impact domains such as healthcare, education, and marketing will ensure that performance-boosting techniques achieve both technical and societal value. These directions provide a roadmap for future research, ensuring that multimedia analytics continues to evolve into a field that produces reliable, interpretable, and impactful knowledge [25] – [50].

4. Conclusion

This study systematically reviewed 26 recent publications in multimedia data analytics, focusing on performance-boosting techniques such as advanced preprocessing, fusion strategies, and deep learning architectures. The synthesis was guided by three research questions.

First, **RQ1** revealed that while multimodal approaches (specifically text–image and audio–video) are now standard, tri-modal and complex cross-modal combinations remain under-explored. Transformer-based models have become the dominant architecture, reflecting a significant shift toward deep attention-based interactions. Second, **RQ2** identified critical barriers to performance, including modality misalignment, noisy data streams, and a lack of standardized benchmarks, which currently limit the scalability of these models. Finally, **RQ3** highlighted future opportunities in lightweight modeling and explainable AI.

To fulfill the roadmap, claim of this study, the following **Actionable Roadmap for Future Research** is proposed:

- **Standardized Benchmarking:** Shift from private or domain-specific datasets to standardized, large-scale benchmarks to allow for objective, reproducible performance comparisons across different fusion architectures.
- **Multi-Level Alignment Evaluation:** Beyond final output accuracy, future work must implement metrics (*Cross-modal Retrieval Recall* atau *Correlation Coefficients*) to evaluate how well individual modalities are synchronized during the fusion process.
- **Robustness & Stress Testing:** Researchers should conduct systematic evaluations under "noisy" conditions—such as missing modality streams or low-quality inputs—to ensure models are reliable for real-world deployment.
- **Efficiency & Hardware Constraints:** Given the increasing complexity of transformers, there is a clear need for "green AI" approaches, focusing on model quantization and pruning to satisfy efficiency constraints for edge-device applications.

In conclusion, multimedia data analytics is a **continuously evolving** research area. While significant progress has been made in deep learning and multimodal integration, its full potential remains undiscovered. By addressing these roadmap priorities, the research community can move toward more accurate, generalizable, and ethically responsible multimedia systems that deliver meaningful societal and commercial impact.

References

[1] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufmann (Elsevier), 2023. ISBN: 978-0-12-811760-6.

[2] E. I. Setiawan, H. Juwiantho, J. Santoso, S. Sumpeno, K. Fujisawa, and M. H. Purnomo, "Multiview sentiment analysis with image-text-concept features of Indonesian social media posts," International Journal of Intelligent Engineering and Systems, vol. 14, no. 2, pp. 521–535, 2021, doi: 10.22266/ijies2021.0430.47.

[3] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," Neural Computation, vol. 32, no. 5, pp. 829–864, May 01, 2020. doi: 10.1162/neco_a_01273.

[4] M. Lymperaiou and G. Stamou, "A survey on knowledge-enhanced multimodal learning," Artificial Intelligence Review, vol. 57, no. 10, Oct. 2024, doi: 10.1007/s10462-024-10825-z.

[5] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, 2017. doi:10.1016/j.inffus.2017.02.003.

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[7] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017, pp. 1103–1114, doi: 10.48550/arXiv.1707.07250.

[8] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep multimodal learning with missing modality: A survey," Oct. 2024. doi: 10.48550/arXiv.2409.07825

[9] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal sentiment analysis: a survey," Jul. 2023, doi:10.1016/j.displa.2023.102563.

[10] L. Che, J. Wang, Y. Zhou, and F. Ma, "Multimodal federated learning: a survey," Sensors, vol. 23, no. 15. Multidisciplinary Digital Publishing Institute (MDPI), Aug. 01, 2023. doi: 10.3390/s23156986.

[11] O. S. Chlapanis, G. Paraskevopoulos, and A. Potamianos, "Adapted multimodal bert with layer-wise fusion for sentiment analysis," Dec. 2022. doi: 10.48550/arXiv.2212.00678

[12] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: a survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 10, pp. 12113–12132, Oct. 2023, doi: 10.1109/TPAMI.2023.3275156.

[13] Y. Wang and M. Wu, "Evaluation of data inconsistency for multimodal sentiment analysis," Jun. 2024. doi: 10.48550/arXiv.2406.03004

[14] K. Kolaski, et al., "Guidance to best tools and practices for systematic reviews," *Systematic Reviews*, vol. 12, no. (issue), 2023, doi: 10.1186/s13643-023-02255-9.

[15] Z. Munn, M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach," BMC Medical Research Methodology, vol. 18, no. 1, Nov. 2018, doi: 10.1186/s12874-018-0611-x.

[16] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, and A. Info, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," World Scientific Annual Review of Artificial Intelligence, vol. 1, Jan. 2023. doi: 10.1016/j.inffus.2022.09.2025

[17] S. Tabakhi, M. N. I. Suvon, P. Ahadian, and H. Lu, "Multimodal learning for multi-omics: a survey," World Scientific Annual Review of Artificial Intelligence, vol. 01, Jan. 2023, doi: 10.1142/s2811032322500047.

[18] D. Moher et al., "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," PLoS Medicine, vol. 6, no. 7. Public Library of Science, Jul. 01, 2009. doi: 10.1371/journal.pmed.1000097.

[19] P. C. Sauer and S. Seuring, "How to conduct systematic literature reviews in management research: a guide in 6 steps and 14 decisions," Review of Managerial Science, vol. 17, no. 5. Springer Science and Business Media Deutschland GmbH, pp. 1899–1933, Jul. 01, 2023. doi: 10.1007/s11846-023-00668-3.

[20] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," Decision Analytics Journal, vol. 3, p. 100073, Jun. 2022, doi: 10.1016/j.dajour.2022.100073.

[21] S. Liang and R. Kusnadi, "Comparative analysis of SVM, XGBoost and neural network on hate speech classification," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 5, pp. 896–903, Oct. 2021, doi: 10.29207/resti.v5i5.3506

[22] A. Abid, S. K. Roy, J. Lees-Marshment, B. L. Dey, S. S. Muhammad, and S. Kumar, "Political social media marketing: A systematic literature review and agenda for future research," Electronic Commerce Research, 2022, doi: 10.1007/s10660-022-09636-7.

[23] W. Zhu, X. Wang, and H. Li, "Multimodal deep analysis for multimedia," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3740–3764, Oct. 2020, doi: 10.1109/TCSVT.2019.2940647.

[24] Y. Christian, T. Wibowo, and M. Lyawati, "Sentiment analysis by using Naïve Bayes classification and support vector machine: Study case Sea Bank," Sinkron: Jurnal dan Penelitian Teknik Informatika, vol. 8, no. 1, pp. 258–274, Jan. 2024, doi: 10.33395/sinkron.v9i1.13141

[25] L. Yan, X. Wu, and Y. Wang, "Student engagement assessment using multimodal deep learning," PLOS ONE, vol. 15, no. 6, Jun. 2020, Art. e0235377, doi: 10.1371/journal.pone.0235377.

[26] F. Khader et al., "Multimodal deep learning for integrating chest radiographs and clinical parameters: A case for transformers," Radiology, vol. 309, no. 1, Oct. 2023, doi: 10.1148/radiol.230806.

[27] D. v. Shah et al., "Building an ICCN multimodal classifier of aggressive political debate style: Towards a computational understanding of candidate performance over time," Communication

Methods and Measures, vol. 18, no. 1, pp. 30–47, 2024, doi: 10.1080/19312458.2023.2227093.

[28] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," Jun. 2019. doi: 10.48550/arXiv.1906.00295

[29] A. Ahmed, K. Shaalan, S. Toral, and Y. Hifny, "A multimodal approach to improve performance evaluation of call center agents," Sensors, vol. 21, no. 8, Apr. 2021, doi: 10.3390/s21082720.

[30] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra- and inter-modality incongruity for multi-modal sarcasm detection," in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021. doi: 10.18653/v1/2020.findings-emnlp.124

[31] W. Jin, L. Zhu, and J. Sun, "Aligning first, then fusing: A novel weakly supervised multimodal violence detection method," Mar. 2025. doi: 10.48550/arXiv.2501.07496

[32] G. Joshi et al., "Multimodal machine learning for deception detection using behavioral and physiological data," Scientific Reports, vol. 15, Dec. 2025, Art. no. 92399. doi: 10.1038/s41598-025-92399-6

[33] A. Fernandez and S. Avinat, "Multimodal sentiment analysis based on video and audio inputs," Procedia Computer Science, Dec. 2024. doi: 10.1016/j.procs.2024.11.082

[34] D. Dimitrov et al., "SemEval-2021 task 6: Detection of persuasion techniques in texts and images," Apr. 2021. doi: 10.48550/arXiv.2105.09284

[35] G. Arya et al., "Multimodal hate speech detection in memes using contrastive language–image pre-training," IEEE Access, vol. 12, pp. 22359–22375, 2024. doi: 10.1109/ACCESS.2024.3361322

[36] A. Alsaeedi, A. AlMansour, and A. Jamal, "Audio-visual multimodal deepfake detection leveraging emotional recognition," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 6, pp. 123–131, 2025, doi: 10.14569/issn.2156-5570.

[37] M. S. AL-Quraishi, S. S. Azhar Ali, M. AL-Qurishi, T. B. Tang, and S. Elferik, "Technologies for detecting and monitoring drivers' states: A systematic review," Heliyon, vol. 10, no. 20. Elsevier Ltd, Oct. 30, 2024. doi: 10.1016/j.heliyon.2024.e39592.

[38] Y. Zhao, M. Komachi, T. Kajiwara, and C. Chu, "Region-attentive multimodal neural machine translation," Neurocomputing, vol. 476, pp. 1–13, Mar. 2022, doi: 10.1016/j.neucom.2021.12.076.

[39] H. Han, Z. Ke, X. Nie, L. Dai, and W. Slamu, "Multimodal fusion with dual-attention based on textual double-embedding networks for rumor detection," Applied Sciences (Switzerland), vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13084886.

[40] D. Bhanushali et al., "LiDAR–camera fusion for 3D object detection," in IS&T Int. Symp. on Electronic Imaging: Autonomous Vehicles and Machines, Jan. 2020. doi: 10.2352/ISSN.2470-1173.2020.16.AVM-257

[41] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," Sensors, vol. 22, no. 16, Aug. 2022, doi: 10.3390/s22165941.

[42] P. Jamadi Khiabani and A. Zubiaga, "Cross-target stance detection: A survey of techniques, datasets, and challenges," Expert Systems with Applications, vol. 283, Jul. 2025, Art. no. 127790. doi: 10.1016/j.eswa.2025.127790

[43] T. Hong, J. A. Choi, K. Lim, and P. Kim, "Enhancing personalized ads using interest category classification of SNS users based on deep neural networks," Sensors (Switzerland), vol. 21, no. 1, pp. 1–17, Jan. 2021, doi: 10.3390/s21010199.

[44] P. P. Tricomi, S. Kumar, M. Conti, and V. S. Subrahmanian, "Climbing the influence tiers on tiktok: A multimodal study," Proceedings of the International AAAI Conference on Web and Social Media, vol. 18, pp. 1503–1516, May 2024, doi: 10.1609/icwsm.v18i1.31405.

[45] Z. Cheng and Y. Li, "Like, comment, and share on TikTok: Exploring the effect of sentiment and second-person view on the user engagement with tiktok news videos," Social Science Computer Review, vol. 42, no. 1, pp. 201–223, Feb. 2024, doi: 10.1177/08944393231178603.

[46] E. Jeong, X. Li, A. Kwon, S. Park, Q. Li, and J. Kim, "A multimodal recommender system using deep learning techniques combining review texts and images," Applied Sciences (Switzerland), vol. 14, no. 20, Oct. 2024, Art. no. 209026. doi: 10.3390/app14209026

[47] J. P. van der Harst and S. Angelopoulos, "Less is more: Engagement with the content of social media

influencers," Journal of Business Research, vol. 181, Aug. 2024, doi: 10.1016/j.jbusres.2024.114746.

[48] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data" Information Processing and Management, vol. 58, no. 5, Sep. 2021, doi: 10.1016/j.ipm.2021.102610.

[49] Y. Yan, H. Fu, and F. Wu, "Multimodal social media fake news detection based on 1D-CCNet attention mechanism," Electronics (Switzerland), vol. 13, no. 18, Sep. 2024, Art. no. 3700. doi: 10.3390/electronics13183700

[50] B. Liang et al., "Multimodal stance detection: New datasets and model," Jun. 2024. doi: 10.48550/arXiv.2402.14298.

[51] Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis Campbell Systematic Reviews, 18, e1230. doi: 10.1002/cl2.1230

.