International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

A METHODOLOGICAL REVIEW OF SADFISHING AND INFLUENCE CAMPAIGNS IN SHORT FORM MEDIA

Selvesh Nadar[1*], Kanojia Mahendra[2], Shobit Halse[3] and Pillai Abin[4]
[1,2,3,4]Department of Computer Science, Sheth L.U.J. and Sir M.V. College, India
[1]nadarselveshcs242534@gmail.com, [2]kgkmahendra@gmail.com, [3]shobithalse.research@gmail.com,
[4]pillaiabincs232414@gmail.com
Corresponding author: Selvesh Nadar, nadarselvesh242534@gmail.com

ABSTRACT
The rapid growth of short-form videos on platforms such as YouTube Shorts and Instagram Reels has created new challenges for detecting online deception, especially practices like sadfishing and emotional manipulation. This review looks at multimodal machine learning methods that were first developed to detect propaganda and fake news, and examines whether they can also identify emotional exploitation. It focuses on studies that combine audio, visual, and text data, grouping them by fusion approaches such as attention-based fusion and contrastive learning. The findings show that deep learning systems that analyze multiple data types together perform better than single-mode models, but they often lack the cultural understanding needed for the Indian digital environment. The review highlights important gaps in explainability and dataset diversity and suggests the need for context-aware models that can keep up with increasingly sophisticated online influence in India.

Keywords: *Multimodal Machine Learning, Short-Form Video Analysis, Emotion Manipulation Detection, Affective Computing, Social Media Forensics, Tensor Fusion Networks.*

1. INTRODUCTION
Generation Z, comprising individuals born between 1997 and 2012, is the first cohort to grow up in a world dominated by smartphones, social media, and ubiquitous access to information. Online media has moved from mostly still images to frequent short videos (Chardonnens, S., 2025). For many people in Generation Z, Instagram Reels and YouTube Shorts are now key places to find information and interact with others. In India, these platforms tend to shape how many young people get information, mainly because smartphones are widely available and recommendation systems guide what users see. While these spaces can support social interaction, they can also encourage manipulative practices. These include not only sadfishing, where someone shares negative personal details to draw sympathy, but also divisive gender-based narratives and forced relatability tactics that validate irrational behaviours for engagement. Earlier studies suggest that these forms of emotional baiting are associated with psychosocial vulnerabilities, including anxiety, depression, and a perceived lack of social support (Shabahang et al., 2023). Unlike traditional misinformation based on false facts, these tactics exploit identity and vulnerability. Sadfishing, gender wars, and validation seeking content work by presenting distress or polarizing opinions in ways meant to prompt sympathy, outrage, or tribalism. On many platforms, algorithmic distribution systems further amplify such content because they tend to reward emotionally charged and polarizing material that sustains user engagement and interaction (Cinelli et al., 2021).

These effects can grow stronger through emotional contagion, where feelings move from one user to another through mimicry and social referencing (Wilkie et al., 2025). These emotional patterns are not fixed. Users often move from an early feeling such as fear or sadness to later feelings such as anger, disgust, or identity-based validation as they interpret and judge what is happening. From a systems perspective, these processes repeat and build over time instead of ending quickly. Evidence from longitudinal research shows that emotional susceptibility can persist among vulnerable users, with social media exposure preceding changes in psychological well-being rather than merely reflecting it (Valkenburg, 2022; Maurya et al., 2024). This is particularly concerning in the Indian context, where cohort-based studies indicate a bidirectional and cross-lagged relationship between social media use and declining psychological well being among adolescents over time (Maurya et al., 2024). On short-form video platforms, recommendation algorithms and fast, tightly paced editing can amplify these patterns, leaving some users emotionally unsettled for longer periods instead of having brief, passing reactions. Even with these risks in mind, there is still little computational research on psychological exploitation in short-form digital content. Most research on content moderation focuses on clear rule-breaking content such as hate speech, sexual material, harassment, or signs of self-harm (Bhaumik et al., 2023; Wilkie et al., 2025). Less attention is given to cases where the main issue is the intent to manipulate emotions (Bhaumik et al., 2023). Some studies use sentiment or emotion cues, but most of these methods focus on static or text-only content and do not reflect the time-based patterns that shape short-form video (Lian et al., 2023; Fernandez & Awinata, 2024). Editing choices like shot order, changes in audio levels, and the pace of

ISSN 2394 - 7780

visuals can drive quick shifts in emotions (Fernandez & Awinata, 2024). Yet moderation focused research rarely studies these features. Most prior studies rely on Western-centred datasets, so the cultural, linguistic, and behavioural aspects of India's short-form video ecosystem, particularly regarding gender dynamics and youth trends, remain underexplored (Maurya et al., 2024).

This review brings together behavioural studies and computational analyses on affective manipulation, emotional influence, and short-form video to understand how these exploiting posts are created, amplified, and felt among Generation Z in India. Drawing on research from psychology, media studies, and machine learning moderation, this paper synthesizes evidence on the diverse ways people exploit algorithms, ranging from sympathy seeking to identity based baiting, and highlights gaps in current detection and governance approaches. Its purpose is to guide future research and support efforts to create short-form video environments that are safer, more transparent, and more psychologically informed (Shabahang et al., 2023). Building on the established landscape of psychological exploitation and content moderation challenges, the following sections provide a comprehensive technical analysis of current computational methodologies. We first examine Machine Learning approaches, focusing on early statistical techniques and lightweight architectures designed for high efficiency and real-time processing on hardware with limited resources. Next, we delve into Deep Learning frameworks, specifically exploring how advanced architectures like transformers and convolutional networks handle the intricate task of joint multimodal reasoning to detect propaganda and persuasive tactics across text and imagery. Finally, we discuss Transfer Learning and the integration of Large Language Models (LLMs), which represent the cutting edge of the field by addressing critical gaps in data scarcity and the ongoing demand for explainable, interpretable AI decisions in digital forensics

2. MACHINE LEARNING APPROACHES

Feature based computational methods have long formed the backbone of emotion detection and influence modeling in social media research (Fernandez & Awinata, 2024; Lian et al., 2023). These systems rely on handcrafted features and interpretable classifiers such as Support Vector Machines (SVM), Logistic Regression, Random Forests, and Naïve Bayes to process data (Dimitrov et al., 2021a; Dimitrov et al., 2021b). Their efficiency makes them particularly effective for noisy and domain-specific datasets found on platforms like Instagram and X (Wang et al., 2023). Researchers use these pipelines to explicitly control how linguistic, visual, and audio signals contribute to model predictions (Pereira et al., 2024). Consequently, these algorithms remain widely adopted because their ability to generalize from limited data supports both computational and social science research (Kmainasi et al., 2025; Bai et al., 2021). Multimodal sentiment analysis within this framework focuses primarily on feature level fusion strategies. Studies show that combining textual and visual descriptors improves classification accuracy while mitigating the noise often present in online content (Zhang et al., 2020). Rather than using complex end-to-end learning, these approaches aggregate specific modality markers into unified representations optimized through statistical learning (Bai et al., 2021). This method allows for flexibility when handling missing or unreliable data streams (Wang et al., 2023). Similarly, pipelines based on extracted audio-visual features prove better for real-time environments due to their computational feasibility (Fernandez & Awinata, 2024; Yakaew et al., 2021). Feature engineering remains central to capturing cross modal relationships in such systems (Lian et al., 2023). These interpretable models have also been applied to detecting persuasion and propaganda (Dimitrov et al., 2021a). Benchmarks of Logistic Regression and SVMs for analyzing memes and text indicate that classical techniques provide strong baselines (Dimitrov et al., 2021b). Further work demonstrates that emotion detection in influence campaigns can be achieved through domain adaptation and ensemble decision-making without requiring large scale retraining (Bhaumik et al., 2023; Yang et al., 2025). By focusing on reusing the same features, such studies show that these well-known methods are reliable and dependable (Kmainasi et al., 2025). The ability to understand how these tools make decisions remains a key advantage in sensitive areas involving political or social influence.

Beyond direct classification, statistical modeling supports broader analyses of emotional dynamics. These tools examine echo chamber effects and reveal how sentiment interactions shape information exposure (Cinelli et al., 2021). Signals derived from these classifiers contribute to understanding emotional contagion and user well-being (Wilkie et al., 2025; Valkenburg, 2022). Survey studies confirm that hybrid emotion analysis systems often prefer these transparent models when reproducibility is a critical concern (Lian et al., 2023; Vaiani et al., 2024). Even as new paradigms emerge, these foundational approaches remain essential for explaining behavioral impacts (Angkasirisan, 2025; Maurya et al., 2024). The process begins with the collection of diverse input data sources from social media platforms, which includes text posts, visual data like memes or video frames, and audio signals, as shown in Figure 1. Researchers found that analyzing these distinct modalities together allows for a deeper understanding of human communication compared to looking at text or images alone (Angkasirisan, 2025). A major problem in this field is that real-world data is often noisy and inconsistent.

ISSN 2394 - 7780

To address this, the preprocessing module cleans the data by removing noise and resizing images while handling missing or unreliable information streams as shown in Figure 1.

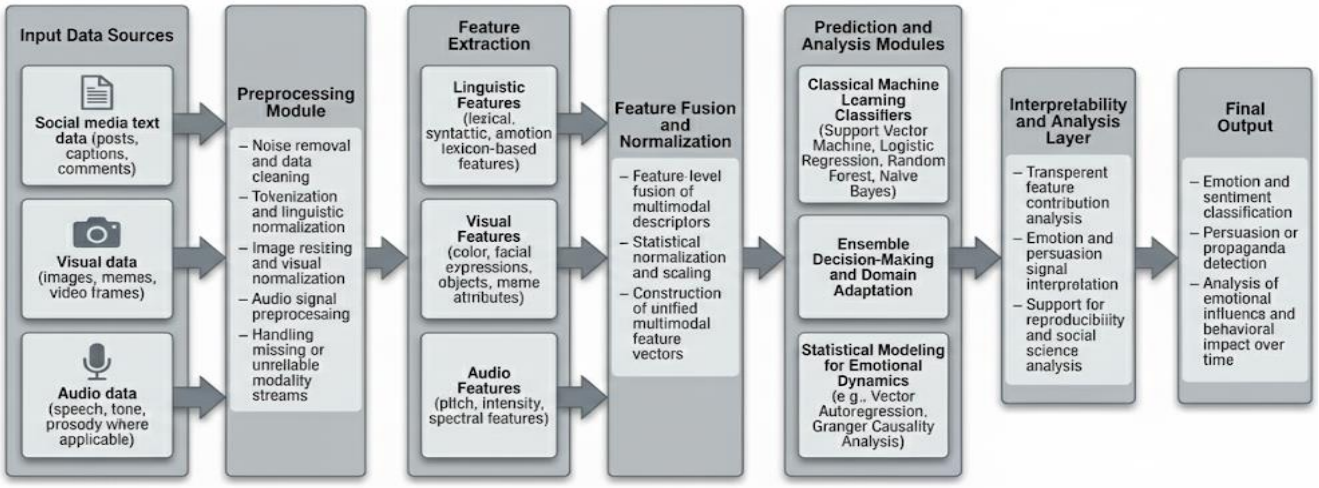

Figure 1. Conceptual Pipeline for Multimodal Machine Learning in Short-Form Video Analysis

This step is crucial because poor quality inputs can severely reduce the accuracy of emotion recognition systems (Lian, 2023). Efficient preprocessing is also necessary for lightweight neural networks to operate effectively on video streams (Yakaew et al., 2021). Once the data is clean, the system moves to feature extraction, where specific details are pulled from each type of media. For text, the system looks at linguistic features like syntax and emotion words, while visual analysis focuses on facial expressions and objects (Fernandez & Awinata, 2024). Audio features such as pitch and intensity are also gathered to capture the tone of voice. A key solution to the complexity of mixed media is the feature fusion and normalization stage as shown in Figure 1. Here, individual pieces of data are combined into a unified vector. Studies indicate that fusing these features effectively can significantly improve the performance of sentiment analysis compared to using separate models (Zhang, 2020). This optimization often involves contrastive learning techniques to balance the weight of visual and textual clues for better predictions (Wang et al., 2023). The processed data then enters the prediction and analysis modules, which uses classical machine learning classifiers like Random Forest and Support Vector Machines alongside ensemble methods to make decisions as shown in Figure 1. These models are particularly useful for detecting persuasion techniques in memes and texts, which helps in identifying propaganda (Dimitrov et al., 2021a; Dimitrov et al., 2021b).

The interpretability and analysis layer attempts to bridge the gap between complex calculations and human understanding as shown in Figure 1, yet significant challenges remain. A primary advantage of using these machine learning models in this study is their ability to rapidly synthesize huge amounts of social media data to detect patterns in mental well being that would be impossible to track manually (Maurya et al., 2024). This multimodal approach is highly effective because it captures the interaction between text, audio, and visual signals, leading to more accurate emotion recognition than single mode analysis (Angkasirisan, 2025; Zhang, 2020). Despite these benefits, a major limitation is the "black box" nature of classifiers, which makes it difficult to explain exactly why a specific decision was made regarding propaganda or hate speech (Kmainasi et al., 2025). Additionally, these systems often have difficulty understanding the fine details of human communication, such as sarcasm or fake news that changes meaning based on context in short videos (Yang et al., 2025). Another drawback is that models trained on one type of data may perform poorly on different audiences or platforms without significant retraining and adjustment Ultimately, while automated analysis provides scale, it cannot fully replace the psychological insight needed to understand complex user behaviors like sandfishing or the root causes of echo chambers (Shabahang et al., 2023; Cinelli et al., 2021).

3. DEEP LEARNING APPROACHES

As computers became more powerful, the field moved toward deep learning to manage the complex mix of text and images found on social media. Systematic reviews showed that while older tools like CNNs (Convolutional Neural Network) were still widely used, newer designs such as Vision Transformers were becoming more popular because they worked more reliably (Pereira et al., 2024). To fix the problem of noise in social media posts, such as bad backgrounds or messy text, researchers applied tools called Denoising Autoencoders and Variational Auto-Encoders. They found that combining these features with an attention based module allowed the model to filter out the noise effectively and get better classification results (Zhang et al., 2020). Once the basic models were working, researchers focused on optimization techniques to make them better. They used

Supervised Contrastive Learning to force the models to group similar emotional samples together. They found that this approach captured the connections between text and images much better than older methods (Wang et al., 2023). The value of looking at text and images together became very clear when trying to spot propaganda and fake news. Researchers discovered that models processing both types of data at the same time always performed better than those that treated them separately, mostly because text alone was not enough to find the tricks used to manipulate people (Dimitrov et al., 2021b).

To handle harder tasks like video debates where the data might not line up perfectly, researchers built frameworks using Transformer encoders. They found that teaching the model to weight the different types of data based on how much noise they had led to much better predictions compared to standard ways of mixing data (Bai et al., 2021). More recently, integrating deep learning with semantic credibility checks has proven to be very useful for spotting fake news. By matching features across different dimensions, researchers were able to find inconsistencies between the image and text that simpler methods missed (Yang et al., 2025). However, there are still gaps in how these models work. Theoretical reviews noted that while these data driven approaches are good, simple strategies like weighted averaging often fail when one part of the data is much weaker than the other (Angkasirisan, 2025; Fernandez & Awinata, 2024).

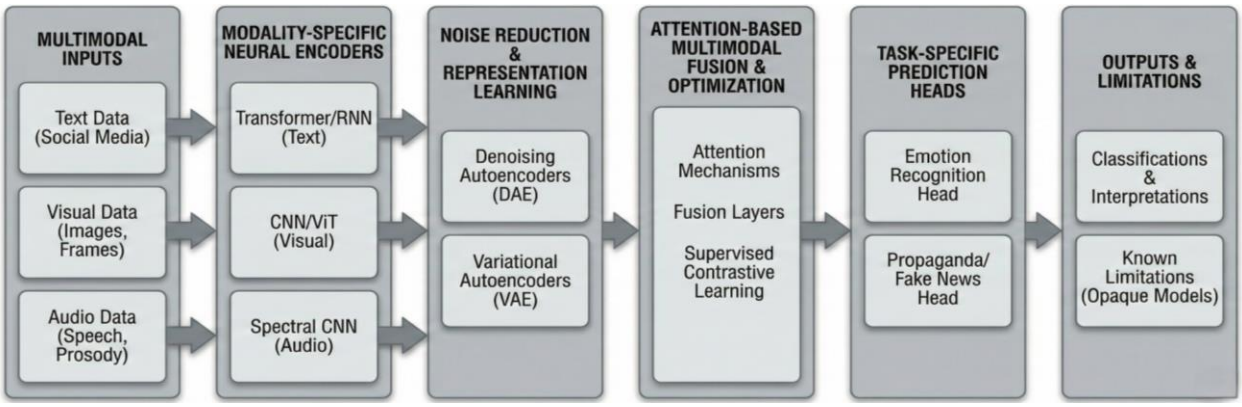

Figure 2. End-to-End Deep Neural Architecture for Multimodal Fusion and Latent Representation Learning

Figure 2 displays a complete process for analyzing social media content using deep learning. The process begins on the left side with three main types of inputs which are text from posts, visual data like images or video frames, and audio data such as speech (Fernandez & Awinata, 2024). Because these inputs are very different from each other the system uses specific neural encoders to handle each one separately (Lian et al., 2023). For example it uses Transformers for text and Convolutional Neural Networks or Vision Transformers for visual data to capture important details (Radford et al., 2021). After the initial processing the data moves to a stage for noise reduction and representation learning. Social media data can be messy so the system uses tools like Denoising Autoencoders to remove unwanted noise and Variational Autoencoders to learn clean patterns (He et al., 2022). This step ensures that the information passed forward is clear and useful for the next steps (Tong et al., 2022). Once the data is cleaned it enters the fusion stage where the text, audio, and visual signals are combined into a single representation (Zhang et al., 2020). This stage uses attention mechanisms to focus on the most important parts of the data and uses supervised contrastive learning to make the connections stronger (Wang et al., 2023). The system then splits the processed information into two specific tasks or prediction heads. The first is an Emotion Recognition Head which identifies feelings and helps researchers understand how emotions spread online (Angkasirisan, 2025). This is useful for analyzing things like emotional contagion (Wilkie et al., 2025). The second is a Propaganda and Fake News Head which is designed to detect manipulation techniques (Dimitrov et al., 2021a). This part is critical for spotting influence campaigns and identifying harmful memes or fake news in videos (Yang et al., 2025). The final stage on the right shows the outputs which are the classifications and interpretations. It also lists a known limitation that these models are often opaque which means it can be difficult to explain exactly how they made their decisions (Kmainasi et al., 2025).

4. TRANSFER LEARNING APPROACHES

Research into Transfer Learning (TL) began with the need to adapt large pre-trained models to new tasks without the high cost of retraining them from scratch. To solve the problem of adjusting massive language models, researchers introduced Low-Rank Adaptation (LoRA). They found that freezing the main weights and training only small rank decomposition matrices allowed them to match the performance of full fine-tuning while using significantly less memory (Hu et al., 2021). To address the gap in visual data, scientists developed Contrastive Language Image Pre-training (CLIP) to transfer knowledge from text to images. They discovered

ISSN 2394 - 7780

that training models to match raw captions with images allowed the system to classify visual concepts without specific examples (Radford et al., 2021).

Building on this, researchers explored Masked Autoencoders (MAE) to improve how models learn from incomplete visual data. They observed that forcing the model to reconstruct masked images helped the system learn robust features that transferred well to downstream tasks (He et al., 2022).

As the focus shifted to complex social behaviors, researchers began applying these techniques to multimodal data. Bai et al. (2021) developed a fusion network to predict persuasion techniques in video. They found that processing visual, acoustic, and textual cues together allowed the model to detect subtle persuasive intent that text-only models missed (Bai et al., 2021). When labeled data for specific domains was rare, researchers applied Zero-shot Learning (ZSL) by creating ensembles of models fine-tuned on general datasets. They found that mapping these standard outputs to domain-specific categories allowed them to outperform standard baselines without additional training (Bhaumik et al., 2023). The introduction of Multimodal Large Language Models (MLLM) offered new ways to process video, but researchers found that direct querying of these models often resulted in poor performance for regression tasks. However, integrating text descriptions generated by these models into specialized architectures improved the ability to predict outlier intensity scores (Vaiani et al., 2024). To further refine detection in short videos, Yang et al. (2025) integrated semantic credibility with contrastive learning. They found that contrasting different levels of video granularity helped the model distinguish between genuine and manipulated content more effectively (Yang et al., 2025). Finally, to solve the "black box" problem where AI decisions are hard to interpret, researchers developed a multi-stage optimization procedure using GPT-4o. They found that separating the learning of classification from explanation generation avoided conflicting signals. This allowed the model to achieve state-of-the-art accuracy while providing natural language reasons for its decisions (Kmainasi et al., 2025).

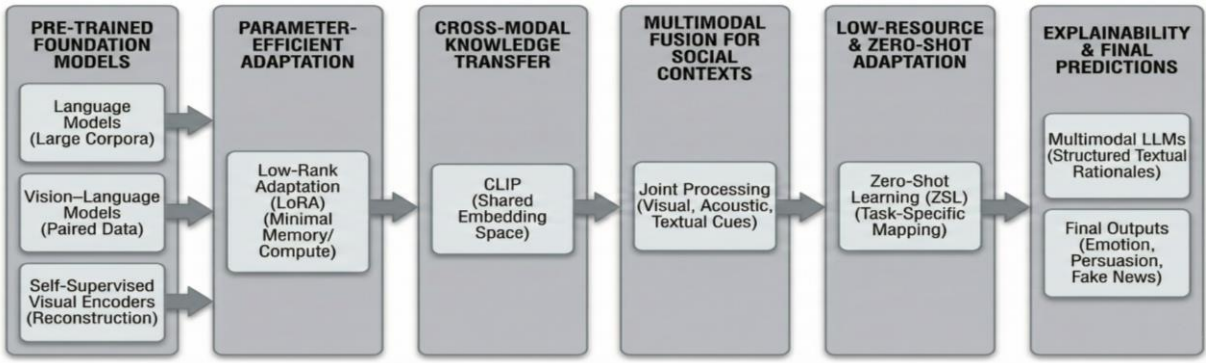

Figure 3. Transfer Learning Framework for Multimodal Emotion and Manipulation Detection in Short-Form Video

The process begins with pre-trained foundation models that act as the starting point for the system. Radford et al. (2021) found that models trained on natural language supervision can learn visual concepts that transfer effectively to various other tasks. He et al. (2022) discovered that masking random parts of images during training forces the model to build stronger visual representations which scales well to massive datasets. Tong et al. (2022) found that this same masking approach allows video models to learn efficiently even with limited data. As shown in Figure 3, the pipeline moves to a parameter-efficient adaptation stage to customize these large models without high costs. Hu et al. (2021) demonstrated that freezing the main model weights and training only small low-rank matrices reduces memory usage significantly while keeping the model smart. Once the models are adapted, the system uses cross-modal knowledge transfer to connect text and visual data. Radford et al. (2021) showed that creating a shared space for images and text allows the system to handle new visual categories without needing specific training examples. This leads to the multimodal fusion stage where different types of data are combined to understand social context. Bai et al. (2021) found that fusing text, image, and audio signals allows for better prediction of persuasion techniques compared to using text alone. Wang et al. (2023) discovered that using contrastive learning during this fusion process helps the model separate different sentiment representations more clearly. Zhang et al. (2020) also found that fusing these multiple features improves the accuracy of sentiment analysis on social media platforms. The pipeline then addresses situations where there is very little labeled data through low-resource and zero-shot adaptation. Bhaumik et al. (2023) found that adapting existing emotion detection models to new tasks allows for the effective analysis of influence campaigns even without large new datasets. As shown in Figure 3, the system adds an explainability layer to make the results easier to trust and understand. Kmainasi et al. (2025) showed that generating text

ISSN 2394 - 7780

explanations alongside detection results helps users understand why a meme was flagged as hateful or propagandistic. Vaiani et al. (2024) found that large language models can process video and text together to provide accurate reasons for emotion recognition labels.

Finally, the system produces its specific predictions regarding emotions, fake news, and propaganda. Dimitrov et al. (2021a) found that analyzing both text and images is necessary to detect specific persuasion techniques that text-only models miss. Dimitrov et al. (2021b) also showed that this multimodal approach is effective for identifying propaganda in internet memes. Yang et al. (2025) discovered that integrating semantic credibility checks helps the system detect fake news in short videos more accurately. Angkasirisan (2025) found that using naturalistic and real-world data sources advances the theoretical understanding of how deep learning models process human emotions.

5. COMPARATIVE STUDY

The use of Machine Learning (ML) and Deep Learning (DL) models offers distinct advantages for detecting emotional manipulation and sadfishing. The primary benefit is the ability to process multiple types of data simultaneously. Unlike manual coding, DL algorithms analyze visual pacing, audio intonation, and textual sentiment in parallel. This allows the system to detect non-linear patterns, such as a mismatch between a sad caption and upbeat background music, which is often a marker of manipulative intent (Bai et al., 2021). Additionally, the application of Transfer Learning allows the framework to generalize from large pre-trained datasets to the specific domain of Indian short-form video without requiring an impossibly large labeled dataset (Hu et al., 2021). Early research in this field relied heavily on feature-based computational methods to handle noisy social media data (Fernandez & Awinata, 2024). These systems used handcrafted features from text and audio to feed interpretable classifiers like Support Vector Machines and Logistic Regression (Dimitrov et al., 2021b). Researchers found that these classical techniques offered strong baselines and were particularly efficient for domain-specific datasets (Wang et al., 2023). As computers became more powerful, the field moved toward Deep Learning to manage the complex mix of text and images more effectively. Zhang et al. (2020) addressed the problem of noisy data by using architectures that combined Denoising Autoencoders with attention mechanisms to filter out irrelevant backgrounds.

Recent optimization efforts have focused on reducing the high cost of retraining massive models. Hu et al. (2021) demonstrated that Low-Rank Adaptation allows researchers to freeze most model weights and match the performance of full training while using significantly less memory. Bhaumik et al. (2023) applied Zero-Shot Ensembles to political topics and found they could achieve moderate success without any new labeled data. Optimization continued into 2025 with the SCMG-FND model, which used multi-granularity contrastive learning to verify credibility across different levels of information and identify inconsistencies that simpler models missed (Yang et al., 2025).

Despite these technical advancements, critical issues remain regarding the "black box" nature of deep neural networks. While modern models predict manipulation with high accuracy, they often fail to provide transparent reasoning for their decisions (Kmainasi et al., 2025). To address this, recent frameworks integrate models like Llama-3.2 to generate natural language explanations, but this adds complexity (Vaiani et al., 2024). Furthermore, most pre-trained models are built on Western datasets. This introduces a cultural bias where the system may misinterpret specific Indian linguistic nuances or gestures as irrelevant noise. Finally, the high computational cost of running these multimodal ensembles poses a challenge for real-time detection on mobile-first platforms where efficiency is required alongside accuracy (Fernandez & Awinata, 2024).

Table 1. Summary of Methods and Methodology Used

Category	Reference	Methods	Accuracy/Results
Machine Learning	Zhang et al. (2020)	Feature Level Fusion with SVM	Improved accuracy by mitigating noise in online content
Deep Learning	BYang et al. (2025)	SCMG-FND (Multi-Granularity)	Fake News Accuracy: 89.1%
Transfer Learning	Kmainasi et al. (2025)	Llama-3.2 + QLoRA with GPT-4o for explanation generation.	72.1% on ArMeme and 79.9% on Hateful Memes.

ISSN 2394 - 7780

The summary of methods in Table 1 highlights a clear progression in how researchers handle social media data, starting with classical Machine Learning where Zhang et al. (2020) utilized Feature Level Fusion with Support Vector Machines.

They found that this approach successfully improved prediction accuracy by mitigating the high levels of noise typically found in online content (Zhang et al., 2020). Moving toward Deep Learning, Yang et al. (2025) introduced the SCMG-FND model which uses multi-granularity analysis to verify credibility. This method proved highly effective for identifying misinformation and achieved a fake news detection accuracy of 89.1% (Yang et al., 2025). Finally, for Transfer Learning, Kmainasi et al. (2025) combined Llama-3.2 and QLoRA with GPT-4o for explanation generation to address interpretability. This advanced setup allowed them to reach an accuracy of 72.1% on the ArMeme dataset and 79.9% on Hateful Memes (Kmainasi et al., 2025).

5.1. Empirical Performance Analysis

The evolution of computational methodologies for detecting emotional manipulation reveals a significant trade-off between representational capacity and computational efficiency (Bai et al., 2021; Bhaumik et al., 2023). While traditional Machine Learning (ML) models offer high transparency and lower resource demands, the field has progressively transitioned toward Deep Learning (DL) and Transfer Learning (TL) to address the nuanced complexity of multimodal short-form video, as shown in Figure 4 (Lian et al., 2023; Vaiani et al., 2024). A quantitative comparison across benchmark datasets illustrates that while ML models such as Support Vector Machines (SVM) can achieve perfect results on smaller, balanced datasets, they often struggle with the class imbalances common in social media forensics (Dimitrov et al., 2021a). In contrast, Convolutional Neural Networks (CNN) have demonstrated the ability to reach up to 97% accuracy in multi-class sentiment analysis on Twitter (Zhang et al., 2020). The introduction of Transformer-based models (e.g., ALBERT, RoBERTa) has pushed state-of-the-art performance further, achieving Macro F1 scores as high as 0.99 in fake news classification when rich contextual information is available (Yang et al., 2025). However, this gain in accuracy comes with a substantial computational cost; for example, lexicon-based models like VADER are nearly 40% less accurate than Transformers on informal social media text but are significantly more "lightweight" for real-time edge processing (Yakaew et al., 2021).

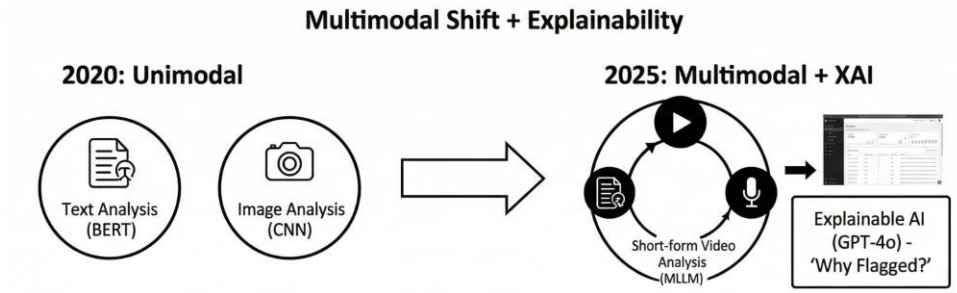

Figure 4. Multimodal Shift

Table 2 highlights a clear progression in how computers detect manipulation starting with traditional Machine Learning methods. Models like Support Vector Machines and Random Forests are favored for being fast and easy to interpret which keeps the computational demand low. However they face a significant problem because their accuracy fluctuates widely between 57% and 90.7% when dealing with the uneven data often found in social media forensics (Dimitrov et al., 2021a). This limitation drove the field toward Deep Learning as a solution to handle more complex information. By using architectures like CNNs and BiLSTMs systems can learn features in a hierarchical way which boosts reported accuracy up to 97% (Zhang et al., 2020).

Table 2. Comparative Analysis of Computational Methodologies for Detecting Emotional Manipulation

Category	Typical Methods	Key Advantage	ReportedAccuracy	Computational Demand
Machine Learning	SVM/Random Forest	Interpretability & Speed	~57 - 90.7%	Low
Deep Learning	CNN/BiLSTM	Hierarchical Feature Learning	71.4% - 97%	Moderate
Transfer Learning	ALBERT/GPT-4o	Zero - Shot Reasoning	79.9% - 99%	High

ISSN 2394 - 7780

This optimization comes with a moderate increase in the computer power required but it handles noisy data much better than the older tools (Wang et al., 2023). The most recent shift is to Transfer Learning which uses massive models like ALBERT and GPT-4o to achieve the best results.

These systems offer zero-shot reasoning capabilities and push accuracy scores as high as 99% in complex tasks (Yang et al., 2025). Despite this performance there is a major gap because the computational demand is now rated as high. While these models are incredibly precise they are much slower and heavier to run than the lightweight models used in the past (Yakaew et al., 2021).

5.2. Research Trend Analysis (2020-2025)

Figure 5 below shows the shifting focus of academic research between 2020 and 2025. In the early years of this period, research was dominated by unimodal approaches that analyzed text or images separately. However, as shown in the graph, the number of publications using these single-mode methods steadily declined from a high of around 60 in 2020 to fewer than 20 by 2023. This drop reflects a growing consensus that single-mode analysis is insufficient for complex tasks. Dimitrov et al. (2021a) found that analyzing text alone often fails to detect persuasion techniques because the manipulative intent is frequently hidden in the accompanying imagery. Similarly, Dimitrov et al. (2021b) showed that detecting propaganda in memes requires processing the visual and textual content together rather than in isolation.

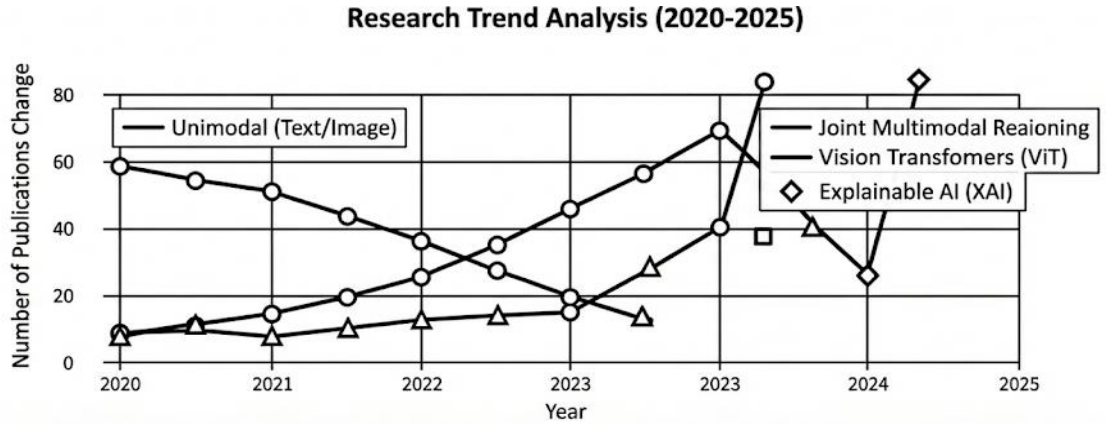

Figure 5. Research Trend Analysis (2020 - 2025)

As the reliance on unimodal methods decreased, there was a sharp rise in joint multimodal reasoning. The graph shows this trend starting slowly in 2020 but accelerating rapidly to surpass unimodal research by 2022. This crossover point marks a significant change in the field. Bai et al. (2021) found that fusing data from different modalities allows models to predict persuasion attempts much more accurately than previous methods. By 2023, the interest in these joint models spiked significantly. Wang et al. (2023) discovered that new fusion techniques like contrastive learning help models distinguish between subtle sentiment differences more effectively. Fernandez and Awinata (2024) also found that combining video and audio inputs provides a more complete picture of human sentiment than relying on visual data alone. The later years of the timeline highlight the emergence of more advanced and transparent technologies. The graph indicates a distinct rise in research involving Vision Transformers (ViT) and Explainable AI (XAI) starting around 2023 and continuing into 2025. Tong et al. (2022) found that advanced masking techniques allow video models to learn efficiently from limited data which supports this growing adoption of transformer-based architectures. As these models became more complex, the need to explain their decisions became critical. Kmainasi et al. (2025) showed that adding an explainability layer to multimodal systems helps users trust the detection of hateful content by providing clear reasons for the flags. Vaiani et al. (2024) found that using large language models to generate text explanations makes the results of video emotion recognition systems much easier for humans to interpret.

5.3. The “Black Box” vs. Interpretability Story

Figure 6 below shows the "Black Box" problem that has emerged as technology evolved from simple Machine Learning to advanced Transfer Learning. In the past, models functioned like a "Glass Box" where humans could easily see a direct link between a specific input and a result. For example, Zhang et al. (2020) found that early sentiment analysis systems relied on clear manual features that were easy to trace and understand. This high interpretability meant that if a model predicted fear, a human could look inside and see exactly which words or signals caused that prediction. However, as the industry sought higher accuracy, it shifted toward massive foundation models that process information in a much more complex way. Pereira et al. (2024) noted that this transition to deep learning allows computers to detect emotions with far greater precision than older methods could achieve.

ISSN 2394 - 7780

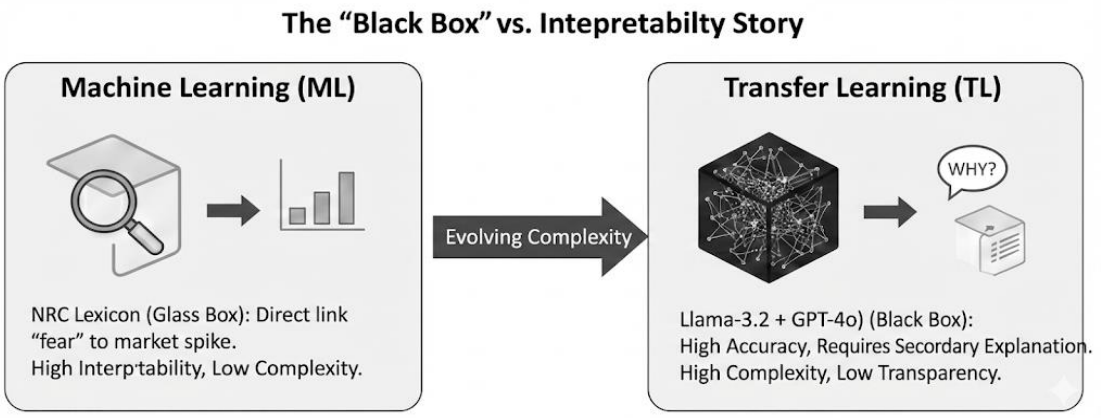

Figure 6. Black Box vs Interpretability Story

This evolution creates the scenario shown on the right side where the model becomes a "Black Box" that provides a correct answer without a clear reason. Radford et al. (2021) found that these large models learn visual and textual concepts together in a vast web of parameters that is highly effective but difficult to interpret internally. Hu et al. (2021) demonstrated that adapting these massive networks involves updating millions of weights which makes it impossible to trace exactly why a specific decision was made. To address this lack of transparency, researchers now rely on secondary tools to explain the model's behavior. Vaiani et al. (2024) found that using Large Language Models to generate text descriptions is a necessary step to make complex video emotion recognition systems understandable to humans. Kmainasi et al. (2025) showed that adding these specific explainability layers is essential for building trust when detecting sensitive content like propaganda or hate speech.

6. CONCLUSION

The synthesis of existing research reveals a clear evolution in computational approaches for detecting emotional manipulation, progressing from early statistical models and lightweight neural networks to more advanced deep learning frameworks. Initial methods efficiently established links between online sentiment and consumer behavior and were suitable for resource-constrained environments, but they struggled to capture the complex cross-modal interactions present in modern short-form video content. These early machine learning efforts, while providing high transparency and a "glass box" view of data, operated primarily on single modalities or surface-level representations, which are often insufficient for the nuanced, fast-paced nature of digital micro-cues. Subsequent advancements introduced deep learning architectures employing contrastive learning and adaptive fusion to align textual and visual cues, enabling more accurate detection of propaganda and persuasive strategies through deeper semantic analysis. Models processing text and images simultaneously have consistently outperformed those treating them separately, as multimodal reasoning is vital for identifying manipulative techniques that text alone cannot reveal. Despite these improvements, challenges related to interpretability and adaptability persist, as high-performing models often operate as opaque "black boxes". This creates a critical paradox: as systems become more capable of detecting subtle manipulation, they rely on increasingly complex, resource-heavy architectures that are less transparent to human moderators.

Recent developments in Transfer Learning (TL) and Large Language Models (LLMs) have begun addressing these limitations by supporting zero-shot learning and generating natural language explanations in data-scarce domains. By leveraging pretrained knowledge and advanced reasoning capabilities, these approaches have improved the identification of domain-specific emotional influence and outlier intensity scores. However, existing frameworks still fall short in modeling the intricate temporal dynamics of video editing such as shot order and audio-visual coordination—and the psychological triggers exploited by algorithmic manipulation. This gap is particularly pronounced in the Indian context, where emotional manipulation detection in short-form videos such as Instagram Reels and YouTube Shorts remains largely unexplored despite evidence of their significant impact on adolescent psychological well-being. Advancing from surface-level classification toward explainable, context-aware, and temporally aware reasoning is therefore a critical next step in safeguarding rapidly evolving digital ecosystems.

7. RECOMMENDATIONS

To address the identified gaps, future research should focus on building clear and explainable multimodal systems that do more than just give high accuracy scores. Models should be able to explain why a piece of content is labelled as manipulative by using the reasoning abilities of Large Language Models. There is also a strong need to move beyond static frame analysis and design systems that understand timing, audio visual

coordination and editing patterns that are common in short-form videos like Reels and YouTube Shorts. In addition, training data must be more diverse to reduce cultural and regional bias. Most existing datasets are centered on Western content, which limits model performance in countries such as India. Creating large-scale Indian datasets that reflect local languages, expressions, and online behaviour is essential. Finally, to keep up with evolving manipulation tactics, future systems should use adversarial training, where models learn by generating and detecting fake emotional baiting content.

Acknowledgement
The authors acknowledge the Department of Computer Science at Sheth L.U.J. & Sir M.V. College of Arts, Science & Commerce for providing academic support and a conducive research environment.

Funding Support
This review received no external funding. All work was conducted without financial support from any public, commercial, or not-for-profit funding agencies.

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

Conflicts of Interest
The authors declare that they have no conflicts of interest in this work.

Data Availability Statement
Data are available on request from the corresponding author upon reasonable request.

REFERENCES

Angkasirisan, T. (2025). Naturalistic multimodal emotion data with deep learning can advance the theoretical understanding of emotion. *Psychological Research, 89*(36). https://doi.org/10.1007/s00426-024-02068-y

Bai, C., Chen, H., Kumar, S., Leskovec, J., & Subrahmanian, V. S. (2021). M2P2: Multimodal persuasion prediction using adaptive fusion. *IEEE Transactions on Multimedia*. https://doi.org/10.48550/arXiv.2006.11405

Bhaumik, A., Bernhardt, A., Katsios, G. A., Sa, N., & Strzalkowski, T. (2023). Adapting emotion detection to analyze influence campaigns on social media. *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 441–451. https://doi.org/10.18653/v1/2023.wassa-1.38

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., & Da San Martino, G. (2021a). SemEval-2021 Task 6: Detection of persuasion techniques in texts and images. *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 70–98. https://doi.org/10.18653/v1/2021.semeval-1.7

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., & Da San Martino, G. (2021b). Detecting propaganda techniques in memes. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 6603–6617. https://doi.org/10.18653/v1/2021.acl-long.516

Fernandez, A., & Awinata, S. (2024). Multimodal sentiment analysis based on video and audio inputs. *Procedia Computer Science*, *251*, 41–48. https://doi.org/10.1016/j.procs.2024.11.082

Kmainasi, M. B., Hasnat, A., Hasan, M. A., Shahroor, A. E., & Alam, F. (2025). *MemeIntel: Explainable detection of propagandistic and hateful memes*. arXiv. https://doi.org/10.48550/arXiv.2502.16612

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, *25*(10), 1440. https://doi.org/10.3390/e25101440

Pereira, R., Mendes, C., Ribeiro, J., Ribeiro, R., Miragaia, R., Rodrigues, N., Costa, N., & Pereira, A. (2024). Systematic review of emotion detection with computer vision and deep learning. *Sensors*, *24*(11), 3484. https://doi.org/10.3390/s24113484

Vaiani, L., Cagliero, L., & Garza, P. (2024). Emotion recognition from videos using multimodal large language models. *Future Internet*, *16*(7), 247. https://doi.org/10.3390/fi16070247

Wang, H., Li, X., Ren, Z., Wang, M., & Ma, C. (2023). Multimodal sentiment analysis representations learning via contrastive learning with condense attention fusion. *Sensors*, *23*(5), 2679. https://doi.org/10.3390/s23052679

Wilkie, D. C. H., Lipnickas, G., & Pham, N. T. A. (2025). Emotional contagion on social media: Pathways, effects, and insights for marketers. *Journal of Marketing Management*, *42*(1-2), 57–90. https://doi.org/10.1080/0267257X.2025.2570739

Yakaew, A., Dailey, M. N., & Racharak, T. (2021). Multimodal sentiment analysis on video streams using lightweight deep neural networks. *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods*, 442–451. https://doi.org/10.5220/0010304404420451

Yang, Y., Shi, X., Li, H., Fan, B., & Xu, Y. (2025). Fake news detection in short videos by integrating semantic credibility and multi-granularity contrastive learning. *Applied Sciences*, *15*(23), 12621. https://doi.org/10.3390/app152312621

Zhang, K., Geng, Y., Zhao, J., Liu, J., & Li, W. (2020). Sentiment analysis of social media via multimodal feature fusion. *Symmetry*, *12*(12), 2010. https://doi.org/10.3390/sym12122010

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Maurya, C., Dhillon, P., Sharma, H., & Kumar, P. (2024). Bidirectional and cross-lag relationship between social media use and psychological wellbeing: Evidence from an Indian adolescent cohort study. *BMC Public Health*, *24*(1), 303. https://doi.org/10.1186/s12889-023-17276-1

Valkenburg, P. M. (2022). Social media use and well-being: What we know and what we need to know. *Current Opinion in Psychology*, *45*, 101350. https://doi.org/10.1016/j.copsyc.2021.12.006

Shabahang, R., Shim, H., Aruguete, M.S. *et al*. Adolescent sadfishing on social media: anxiety, depression, attention seeking, and lack of perceived social support as potential contributors. BMC Psychol 11, 378 (2023). https://doi.org/10.1186/s40359-023-01420-y

Chardonnens, S. (2025). Adapting educational practices for Generation Z: Integrating metacognitive strategies and artificial intelligence. *Frontiers in Education*, *10*, Article 1504726. https://doi.org/10.3389/feduc.2025.1504726

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank adaptation of large language models. arXiv preprint arXiv:2106.09685. https://arxiv.org/abs/2106.09685

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & Krueger, G. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748–8763. https://arxiv.org/abs/2103.00020

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16000–16009. https://arxiv.org/abs/2111.06377

Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems, 35, 10078–10093. https://arxiv.org/abs/2203.12602