IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, July 2026
Impact Factor: 8.2

Systematic Review of the Architectural Transition in Sentiment Analysis: From Hybrid Deep Learning to Generative Large Language Models

Khushi Gautam[1] and Rupali Bhartiya[2]
M.Tech Scholar, Department of Computer Science Engineering[1]
Head of Department, Department of Information Security and Cloud Computing [2]
Shri Vaishnav Institute of Information Technology, Indore[1]
Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, M.P., India[2]
khushigautam2208@gmail.com[1], rupalibhartiya@svvv.edu.in[2]

Abstract : *The rapid evolution of natural language processing has driven sentiment analysis from simple lexicon based tools to sophisticated generative large language models. This systematic review charts that architectural transition, covering three distinct eras: rule based and statistical machine learning, hybrid deep learning architectures (notably CNN LSTM), and the current paradigm of transformer based and generative LLMs. Following PRISMA and Kitchenham guidelines, we synthesise findings from high impact publications (2020–2025) across IEEE, Elsevier, Springer, and ACL. The performance measures, contextual reasoning abilities, and other challenges like model variability, sarcasm detection, multimodal fusion, and interpretability form part of our analysis. Additionally, issues surrounding sustainability are addressed via Green AI techniques such as quantization and knowledge distillation. The findings show that although discriminative fine-tuned models continue to perform excellently in narrow classification tasks, generative LLMs possess impressive zero shot reasoning and flexibility but have issues with inconsistency and lack of transparency. We conclude by identifying key research gaps – deterministic benchmarking, uncertainty aware calibration, autonomous multimodal reasoning, and agentic explainability – and propose a roadmap for future work. This survey serves as a comprehensive resource for researchers and practitioners navigating the shifting landscape of sentiment analysis.*

Keywords: Sentiment analysis; opinion mining; large language models (LLMs); hybrid deep learning; transformer architectures; zero shot learning; model variability; explainable AI (XAI); Green AI; multimodal sentiment analysis.

I. INTRODUCTION

Sentiment analysis – also known as opinion mining – has become a cornerstone of computational linguistics and social data science. In an age where digital platforms generate exabytes of unstructured textual content daily, the ability to automatically discern human emotions, opinions, and stances is no longer a luxury but a strategic necessity. From e commerce giants fine tuning product recommendations to financial institutions gauging market sentiment, and from public health agencies monitoring vaccine hesitancy to governments tracking citizen feedback, the applications are vast and high stakes.

The field has undergone a remarkable metamorphosis over the past decade. Early systems relied on hand crafted lexicons and shallow machine learning classifiers that treated sentences as bags of words – fast and interpretable, but brittle in the face of sarcasm, negation, or context dependent meaning. The advent of deep learning brought hybrid neural architectures, such as CNN LSTM and attention based models, which learned hierarchical features and sequential dependencies, pushing accuracy beyond 90% on benchmark datasets. More recently, the transformer

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

revolution – epitomised by BERT, RoBERTa, and their derivatives – introduced bidirectional context and self attention, effectively capturing long range relationships and setting new state of the art records. Today, we are witnessing yet another seismic shift: generative large language models (LLMs) like GPT 4, Llama 3, and DeepSeek R1 are redefining sentiment analysis from a discriminative classification task to a generative reasoning problem, enabling zero shot and few shot inference with human like justifications.

However, this rapid architectural progression is not without its growing pains. The very factors that make LLMs powerful – their scale, stochasticity, and prompt sensitivity – also introduce model variability, inconsistency across providers, and opacity, raising concerns about trustworthiness in critical domains. Moreover, the environmental cost of training and deploying these behemoths has spurred a "Green AI" movement, urging the community to treat energy efficiency as a first class evaluation metric.

This systematic review aims to provide a comprehensive, up to date mapping of the architectural transition in sentiment analysis, from hybrid deep learning to generative LLMs, while critically examining the trade offs involved. Specifically, our work is guided by the following research questions:

1. **RQ1:** What are the key architectural milestones and their underlying mechanisms in the evolution of sentiment analysis?

2. **RQ2:** How do hybrid deep learning models (CNN LSTM, attention based) compare with transformer based and generative LLMs in terms of accuracy, generalisation, and computational cost?

3. **RQ3:** How can the emerging challenges – especially model variability, interpretability, and sustainability – be addressed?

4. **RQ4:** Which of the possible future directions are the most promising ones for reliable, efficient, and interpretable sentiment analysis?

   In order to address these research questions, we adopted a systematic literature review process consistent with PRISMA and Kitchenham procedures, which entailed examining peer reviewed journals and preprints between 2020 and 2025 from sources such as IEEE Xplore, ScienceDirect, ACM Digital Library, and ACL Anthology.

The main contributions of this review are:

1. A structured chronology of sentiment analysis architectures, categorised into statistical, hybrid deep-learning, transformer-based, and generative LLM paradigms.
2. A comparative analysis of performance metrics, strengths, and limitations for each paradigm, supported by recent empirical studies.
3. An in-depth discussion of critical challenges: model variability, sarcasm/irony detection, multimodal fusion, explainability, and environmental sustainability.
4. A forward-looking perspective that identifies open problems and proposes a research agenda for the next generation of sentiment analysis systems.

The rest of this paper will be divided as follows. Section II will provide additional details regarding the methodological approach for the systematic review. Section III provides information about the history of architectures in sentiment analysis, from lexicon-based and statistical models to the hybrid deep learning approach (CNN-LSTM), Transformer architecture, and specific approaches such as SiEBERT. Section IV will address the generation paradigm, including zero-shot prompting and instruction-tuned inference, along with other issues that can be addressed in the context of

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

model variability. Section V will cover new challenges that arise in the context of sentiment analysis research, including complex linguistic constructions such as sarcasm and irony, multimodal sentiment analysis, and fusion methods. Section VI will include information on the topics of sustainability, model compression, and Green AI. Section VII covers the issue of explainability, specifically the explainability vs performance trade-off in sentiment models. Section VIII includes domain-specific benchmarking, financial applications, medical applications, educational applications, social governance, and others.

II. SYSTEMATIC REVIEW METHODOLOGY AND LITERATURE SYNTHESIS

An easy We followed standard rules for reviewing research to keep things thorough and avoid repeats in this fast-moving field of natural language processing. The incorporation of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines provides a systematic framework for the selection and analysis of primary research articles published between 2020 and 2025. Moreover, the implementation of Kitchenham's rules for systematic literature reviews in computer science ensures that the research questions about method-related progress and application domain are addressed via a transparent and verifiable process.

We searched top academic databases like IEEE Xplore, ScienceDirect, ACM Digital Library, and ACL Anthology, and "identified a large number of papers showing just how fast this field is moving.

Recent meta-analyses have reviewed as many as 4,709 papers from reputable sources such as Springer and Elsevier, eventually narrowing the selection to studies that provide the most significant technical contributions to deep learning and generative architectures.

TABLE I
IMPLEMENTATION OF THE PRISMA PROTOCOL (2020–2025)

Step in SLR Protocol	Implementation Description	Objective for Modern Sentiment Analysis
PICOC Definition	Identification of Population (Reviews, Social Media), Intervention (LLMs, Hybrids), Comparison (Lexicon Baselines), and Outcome (F1-score, Precision)	Delineate the scope of the architectural transition from discriminative to generative models
Reputable Source Selection	Inclusion of IEEE (26 articles), Elsevier (24), and Springer (19) based on 2024 citation data	Ensure comprehensive coverage of high-impact AI research
Search Refinement	Iterative calibration of query strings focusing on "Transformers," "Hybrid CNN-LSTM," and "Zero-shot LLMs"	Filter for technical depth and methodological novelty in recent SA architectures
Quality Assessment	Application of numerical scales to evaluate dataset quality, feature selection, and classifier choice	Quantify the reliability of reported accuracy and F1-score benchmarks
Data Synthesis	Categorization into statistical, deep learning, and generative paradigms	Map the chronological and technological evolution of the field

III. EVOLUTION OF SENTIMENT ANALYSIS ARCHITECTURES

A. The Foundations of Affective Computing: From Lexicons to Statistical Models

The early development of sentiment analysis relied on the surface properties of the language used. Lexicon-based approaches such as VADER (Valence Aware Dictionary and sEntiment Reasoner) utilized predefined dictionaries where each word had sentiment values that represented polarity and intensity

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

Lexicons remain effective in performing fast sentiment classification for small-size datasets because of their simplicity and high speed of execution. The final sentiment score is the sum of the emotional scores of the individual words used in it. But the biggest limitation of lexicon-based models is that they cannot cope with the linguistic complexity of a sentence including such nuances as irony or negation.

Data-driven Machine Learning approaches SVM, Naïve Bayes and logistic Regression signified the emergence of approaches that could automatically find relevant features. Vector space based models, such as Bag-of-Words and TF-IDF (Term Frequency-Inverse Document Frequency) categorized sentences according to their sentiment scores. Although these approaches produced excellent results in terms of accuracy: usually exceeding 80% for structured data sets such as IMDB — they could not account for the sequential dependencies and context in between words in the sentence.

B. The Era of Hybrid Deep Learning: Spatial-Temporal Feature Fusion

The move from statistical models to neural networks became necessary because of the nature of non-linearity and complexity associated with natural languages. Deep learning models offered the possibility of carrying out feature extraction and classification through one process, outperforming the conventional approaches in tasks where deep comprehension of user-created data was needed.

a) Architecture of the CNN-LSTM Hybrid Framework

Among the major advances in the recent period is the creation of hybrid models that combine CNN with RNN models especially LSTMs.

In this case, the CNN layer serves as an important feature extractor where convolutional kernels detect the local n-gram relationship and spatial pattern in the text.This would enable the model to learn essential sentences, such as specific adjectives or phrases, regardless of their position within the sentence. Generally, max-pooling layers follow convolution to highlight important features and reduce complexity.

Yet, CNNs are naturally restricted by the locality of their actions, that hinders efficient modeling of long-distance dependencies. To solve this, the feature representation produced by CNN outputs are forwarded to an LSTM layer which is designed to retain long-term dependencies through a complex gating mechanism. By adding bidirectional LSTMs (Bi-LSTM) , the performance is significantly boosted by allowing the model to capture context from both previous and next tokens, which are crucial for understanding sentences where the sentiment is influenced by subsequent clauses.

Three gates that form the math behind the gates of memory cells in an LSTM are:

Input Gate (i_t): Indicates what new information to be added in the cell state.

$i_t = \sigma(W_i \cdot X_t + U_i \cdot h_{t-1} + b_i)$

Forget Gate (f_t): Indicates the information being carried forward from the last state.

$f_t = \sigma(W_f \cdot X_t + U_f \cdot h_{t-1} + b_f)$

Output Gate (O_t): Helps in controlling the information added in the hidden state from the cell state.

$O_t = \sigma(W_o \cdot X_t + U_o \cdot h_{t-1} + b_o + V_o \cdot C_t)$

Using these layers in combination allows the hybrid models to learn about contextual features, as well as retain the sequence information. The recent use of hybrid models in 2025 indicates that the models have an accuracy rate between 91% to 98.9% on various datasets including IMDB, Amazon review and educational performance dataset.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, July 2026
Impact Factor: 8.2

TABLE II
Function and Benefits of Hybrid Deep Learning Components

Hybrid Component	Function in Sentiment Analysis	Key Benefit and Metric Improvement
Embedding Layer	Pretrained vectors (GloVe, Word2Vec)	Captures semantic similarity between words
CNN Layers	Spatial feature extraction (n-grams)	Identifies key sentiment-bearing phrases with 91% accuracy
LSTM Layers	Preserving long-term dependencies	Captures semantic flow and context in longer reviews
Attention Module	Dynamic weighting of salient tokens	Focus aspect-relevant terms, achieves F1-scores up to 0.98
Max Pooling Layer	Structural organization and reduction	Organizes the sequence structure for global classification

C) The Transformer Revolution: Bidirectional Context and SiEBERT

The introduction of the Transformer architecture in 2017 marked a significant shift that has reduced recurrent architectures by replacing them with self-attention mechanisms. Transformers permit parallel data processing and a thorough understanding of context and it set new standards for sentiment classification accuracy and setting the current "System 1" processing baseline for sentiment tasks.

a) Contextual Embedding Excellence: BERT and RoBERTa

Models such as BERT (Bidirectional Encoder Representations from Transformers) modified the field by providing better contextual representations. The architecture of BERT utilizes a Masked Language Model (MLM) objective, where the model predicts arbitrarily masked tokens within an input sequence, hence learning the relationships between words in both direction at the same time. RoBERTa (Robustly Optimized BERT Pretraining Approach) further improved this by introducing dynamic masking, extensive training data and larger batch sizes and hence establishing itself as a highly consistent foundation for sentiment tasks.

A significant development in this series is the SiEBERT (Sentiment in English) model. It is a fine-tuned checkpoint of RoBERTa-large. Unlike models trained on single datasets like SST-2, SiEBERT was fine-tuned on a meta-corpus of 15 different datasets including- movie reviews, tweets and product titles. This approach assures high generalizability, with the model attaining an average test set accuracy of 93.2%. Robustness is checked using leave-one-out evaluation technique which shows only 3% performance decrease when evaluated on unseen text types, highlighting the effectiveness of multi-domain transfer learning in creating a general-purpose sentiment engine.

b) Self-Attention and the Interpretable Gradient

The self-attention mechanism at the heart of Transformers calculates the correlation between tokens by transforming input embeddings into Query (Q), Key (K), and Value (V) matrices. The importance of a token to the current context is determined by the dot product of its Key with the Query of other tokens, normalized via a softmax function :

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

This process provides a form of build-in interpretability, as researchers can visualize attention weights to identify which words the model focused on for a particular prediction. However, while attention weights correlate with sentiment-laden terms, they are usually inadequate for complete interpretability, as they may capture statistical correlations rather than logical connections which leads to the development of specialized explainability frameworks like DeepBERT-XAI.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, July 2026
Impact Factor: 8.2

TABLE III
Comparative Analysis of Transformer-Based Models for Sentiment Analysis

Model Variant	Model Variant	Evaluated Accuracy (Avg.)	Primary Advantage
DistilBERT	Knowledge Distillation	78.1% (Generalization)	40% smaller, 60% faster than BERT
BERT-base	Pretrained Bidirectional	~90% (Structured)	Foundation for transfer learning
RoBERTa-base	Optimized Training	94-95% (Stable)	Dynamic masking and larger data
SiEBERT	Meta-corpus Fine-Tuning	93.2% (Meta-SA)	Outperforms single-domain models by 15%
FinBERT	Domain-specific Tuning	93.27% (Financial)	Captures nuanced financial tone

IV. THE GENERATIVE PARADIGM: ZERO-SHOT REASONING AND PROMPT-BASED INFERENCE

The most recent architectural transition is the shift from discriminative classification models to generative Large Language Models (LLMs). While earlier neural models focused on predicting a class label, generative LLMs like GPT-4, Llama 3, and DeepSeek-R1 function as autoregressive next-token predictors that can explain their reasoning in human-understandable terms.

A) Instruction-Tuned Inference and Few-Shot Learning

In the LLM era, sentiment analysis is increasingly performed using zero-shot or few-shot prompting, which eliminates the need for expensive domain-specific fine-tuning. This flexibility enables domain experts to modify the behavior of the model immediately using natural language commands, updating the sentiment parameters based on changing market language or changing risk tolerance levels.

Experimental results demonstrate that while fine-tuned encoders still maintain an edge in niche classification—especially as the number of labels increases—zero-shot LLMs perform remarkably well for binary sentiment classification. For example, GPT-4 and Aya-Expanse-8B have been found to outperform classifiers trained on full human-labeled datasets in specific sentiment tasks. However, for more complex topic classification (e.g., intent or scenario), custom classifiers trained on full data still maintain a significant performance lead.

B) The Model Variability Problem and Temperature Scaling

A serious issue which was introduced by the shift to generative architectures is the "Model Variability Problem" (MVP) which is characterized by inconsistent sentiment outputs for the same input in multiple trials. This variability arises from the stochastic nature of probabilistic text generation and the high sensitivity of these models to input wording and decoding parameters.

Randomness in LLMs is largely controlled by the temperature (T) hyperparameter, which shapes the probability distribution created by the softmax function before token sampling :

$$P(x_i) = \frac{\exp(z_i/T)}{\Sigma_j \exp(z_j/T)}$$

1. **Low Temperature (T ≈ 0.1 – 0.5)**: The distribution will become very "peaky" around the maximum probable token thus making the output deterministic and predictable, which is desirable for sentiment analysis tasks.

2. **Moderate Temperature (T ≈ 0.7 – 1.0)**: It keeps a balance between coherence and diversity, reflecting the unchanged output of the training distribution.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, July 2026
Impact Factor: 8.2

3. **High Temperature (T > 1.0)**: The distribution will flatten, resulting in a higher probability of picking up low-probability tokens. This 'creativity' can sometimes help, but more likely will cause hallucinations and quality degradation in evaluation tasks.

C) Performance Disparities and Language Bias

Inter-model consistency remains a significant concern. In a comparative study of recent LLMs, the Intra-class Correlation Coefficient was calculated at 0.061 for identical student feedback which indicates a near-total disagreement between different models. This suggests that machine-generated sentiment from LLMs holds hidden emotional tone that changes consistently by provider, making model calibration and error analysis essential for trustworthy deployment in high-risk sectors.

TABLE II
Sources of Model Variability in Generative Large Language Models

Source of Variance	Mechanism of Influence	Impact on Sentiment Analysis Output
Temperature Scaling	Flattens token probability distribution	Increases output randomness and inconsistency
Prompt Sensitivity	Minor phrasing shifts in instruction	Causes significant polarity flips and sentiment propagation
Context Length	Retrieval of long-range dependencies	Accuracy falls below 10% after 10 turns in zero-shot settings
Structured Output Errors	Next-token prediction of explanations	LLMs may generate rationales instead of simple labels (up to 10% error)

V. EMERGING CHALLENGES: SARCASM, IRONY, AND MULTIMODALITY

A) Complex Linguistic Structures: Sarcasm and Irony Detection

One of the key theoretical challenges in terms of the architecture change is related to the problem of recognizing implicit sentiments, like sarcasm and irony when the meaning contradicts the literal words. Many opponents criticize the large language models, comparing them to parrots due to their lack of true understanding of the language and inability to recognize such concepts.

a) Pragmatic Metacognitive Prompting (PMP)

PMP is an emerging technique used in understanding irony in language models. The theory involves balancing the available data and making sense of the context of the statement by detecting any clue from the sentence. The following are the various types of models used in this kind of reasoning:

TSPM - This model stresses how people utilize both literal and non-literal meaning in communication and how they are related to context.Theory of Pretense in Irony - This theory explains how a person can communicate something that goes against his/her beliefs.Echoic Reminder Theory - This theory explains how a person can make reference to something previously said by him or her to bring home the absurdity of the matter.

PMP has shown that researchers are able to enhance the performance of language models like GPT-40-mini and Gemini-1.5-flash by creating significant improvements (up to 46%) in the model's ability to detect irony when there are both literal and non-literal meanings present in the same sentence.

b) Reinforcement Learning for System 2 Reasoning

While traditional LLMs rely on "System 1" rapid processing, sarcasm detection is increasingly viewed as a "System 2" task requiring slow, deliberate, multi-step reasoning. New approaches decompose sarcasm detection into three dimensions: language, context, and emotion. To reach optimal levels of performance on the task, researchers have

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

applied reinforcement learning (RL) using customized reward models.These reward models guide the model to focus on rhetorical devices (e.g., hyperbole, punctuation) and emotional inconsistencies during the reasoning process.

B) Multimodal Sentiment Analysis (MSA) and Fusion Strategies

The exponential growth of multimedia content on social media has catalyzed the move toward Multimodal Sentiment Analysis (MSA), where textual data is fused with visual, auditory, and even physiological signals. Traditional sentiment analysis restricted the capability to interpret subtle emotional expressions that are often clarified by tone of voice or facial expressions.

a) Technical Fusion Mechanisms

Current methodologies are grouped by how they combine different data types:

Early and Late Fusion: Some approaches merge data right from the start, while others wait until the end to make a decision based on separate analyses.

Hybrid and Tensor Fusion: Other researchers have been doing similar types of work by using combinations or usages depending upon data sources and how they will work together.Attention-Based Fusion: Other researchers have developed ways to use attention to help determine what information is relevant and how different types of data relate to that information such as through the implementation of the A-MSDA framework.Recently, researchers have been using big multimodal models like GPT-4o and LLaMA 3 to bring in extra knowledge and reasoning skills. These models perform well, but they are computationally expensive, so people are looking into ways to make them more efficient.

b) Reasoning Distillation in MSA

In order to meet the needs of resources of the MLLMs, the "MulCoT-RD" framework utilizes an intermediate-size open-sourced model as the assistant to create high-quality data to provide soft label distillation signals. With such multi-task learning strategy, the light-weighted 3B-parameter MLLM is able to deliver better sentiment classification results in a very high level of interpretability.

TABLE III

Multimodal Sentiment Analysis: Datasets and Primary Challenges

MSA Dataset/Feature	Modalities Involved	Primary Challenge Addressed
MuSe-CaR	Audio, Visual, Textual	Trustworthiness and engagement recognition
19th-Century Newspapers	Historical Text, Metadata	Irony detection in specialized historical contexts
Physiological Signals	Wearable sensor data	Detecting emotions even when other modalities conceal them
Contrastive HCRL	Heterogeneous Semantic Features	Modeling cross-modal interactions via consistency learning

VI. SUSTAINABILITY AND THE TRANSITION TO GREEN AI

The rapid development of the sentiment analysis models leads us into the age of "Red AI" where SOTA accuracy is achieved with exponentially increasing model parameters and energy consumption. Thus, the "Green AI" paradigm raises that the energy efficiency and carbon-awareness should be regarded as first-class evaluation criteria.

A) Quantization and Model Compression

This is one of the best ways to lower both LLM expenses and environmental damage.

1. Four Bit (INT4) Quantization - Reduces the model size by four to eight times, meaning that large models can be used on standard hardware with little to no degradation in performance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

2. Six Bit (FlexQ) Quantization - Will provide a new six bit weight with adaptively stored eight bit activation layer as of 2025 in order to provide the best compromise between compactness and performance.

3. Model Pruning - Is a method of removing unneeded weight from a neural network, thereby creating a smaller and faster machine learning system.

B) Knowledge Distillation Frameworks (KNOWDIST and CompEffDist)

The use of knowledge distillation (KD) to improve sentiment analysis has progressed from simple transfer of soft-labels to more sophisticated forms of frameworks.

1. **KNOWDIST:** This is a two-stage approach of distilling knowledge which splits the knowledge distillation process into the knowledge portion and the alignment portion. The approach utilizes a multi-perspective prompting technique that targets expression, target, and emotion (A summary of KD can be found in Section 1 of the introduction).

2. **Comp Eff Dist:** Introduces attribute-based automatic instruction construction and difficulty-based data filtering. This method works like the usual methods but uses less than 10% of the training data. Thus, this greatly minimizes environmental damage.

TABLE IV
Green AI Strategies for Sustainable Sentiment Analysis

Green AI Strategy	Mechanism of Implementation	Energy/Cost Benefit
DistilBERT	Distillation of BERT capacity	40% size reduction, 60% faster inference
INT4 Quantization	Conversion from FP32 to 8-bit/4-bit	4x to 8x memory savings
HEQ (Custom Multipliers)	Hardware-aware quantization	Up to 4000x power reduction
CompEffDist Filtering	Difficulty-based data reduction	performance parity with 90% less data

VII. EXPLAINABILITY AND THE INTERPRETABILITY TRADE-OFF

Explaining AI (XAI) is shifting the paradigm of complex sentiment models which cannot be explained anymore since they are losing their transparency due to obscured transparency. Lack of transparency leads to problems regarding fairness and loss of trust. With explainability, we will be able to gain insights on how complex models function using the concept of XAI.

A) Explainable AI Taxonomy:

Classifying Post-Hoc vs. Intrinsic Explainability Interpretability is classified based on timing and mechanism of explanations:

1. **Intrinsic Explainability**: Models that are inherently interpretable based on their simple nature (i.e., Decision Trees, Linear Regression) or that contain embedded attention mechanisms. As a result, they are most often used in high-stakes areas (i.e., medical and financial domains) where a large degree of transparency is mandated by regulation.

2. **Post-Hoc Explainability**:Different methods of making sense of complicated "black box" models that have already been constructed. Two of the more common approaches to post hoc interpretability include SHAP (Shapley Additive Explanation) and LIME (Local Interpretable Model-Agnostic Explanations). In many cases, the use of SHAP is preferred to that of LIME because SHAP has both mathematical guarantees and stable

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, July 2026
Impact Factor: 8.2

outputs. Future of AI - Agentic AI & Neuro-Symbolic Agents In 2025, AI will progress toward being "agentic" or capable of explaining themselves actively. Agent57 of DeepMind is one of the examples of agents that combine deep neural networks and knowledge graphs, utilizing both pattern recognition and logical inference for self-explanation.

B) Neuro-Symbolic Agents and ROI

Looking ahead into 2025, there is a tendency of AI to evolve towards "Agentic AI" – agents capable of explaining themselves. Agents such as Agent57 of DeepMind which combine neural networks and knowledge graphs, pattern recognition and logical inference can lead to 30% increase in ROI compared to black boxes. Moreover, studies showed that XAI can lower physician override rate in medical diagnostics by 19% and increasing accepting rate of financial advice by 41%.

VIII. DOMAIN-SPECIFIC PERFORMANCE BENCHMARKS AND APPLICATION TRENDS

The shift in architecture has played out differently across high-stakes fields, with specialized models often beating general ones.

Why Do Target-Based Financial Sentiment Analysis (TBFSA)? The role of sentiment analysis within the financial market is crucial because it helps to correctly control one's portfolio and make quick and rational investment decisions. Until recently, the main sentiment analysis models were FinBERT and DistilfinRoBERTa, which operated successfully up until 2025; yet, generative LLMs, including DeepSeek-R1 and ChatGPT-o1, surpassed all of these models in the benchmark testing without any substantial fine-tuning.

A) Public Health and Clinical Workflows

Sentiment analysis in public health helps understand patients' opinions about vaccination, healthcare reforms, and the quality of services offered. According to scientific research, there has been a growing use of advanced models and LLMs in discussing health issues on a large scale, yet data quality, language complexity, and privacy regulations (HIPPA/GDPR) remain significant obstacles. While LLMs surpass basic models like VADER, the latter require specific prompts with contextual information to provide clinically accurate predictions.

TABLE VII

Domain-Specific Performance Benchmarks and Emerging Trends (2025–2026)

Domain	Leading Model Architecture	Performance Metric	Emerging Trend (2025–2026)
Finance	DeepSeek-R1 / ChatGPT-o1	93.27% F1-score	Zero-shot target-level analysis
Healthcare	BERT-based fine-tuned models	High accuracy on vaccination	Mental health surveillance via LLMs
Education	Hybrid CNN-LSTM / BERT	98.93% Accuracy	Academic performance and course review mining
Social Care	XAI-augmented Transformers	F1-score of 71.04%	Interpretable sentiment for fan engagement
Infrastructure	SVM / Naive Bayes	Baseline Performance	Public policy sentiment monitoring

XI. CONCLUSION

Sentiment Analysis has undergone a great change between 2015 and 2026; essentially how computers assess what people mean has changed radically due to the improvement in technology by transitioning from traditional lexicon-based tools (which rely entirely on predefined rules) to hybrid deep learning models (which apply deep learning methods to large volumes of data). Furthermore, the advent of Transformer models introduced a global bidirectional

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, July 2026
Impact Factor: 8.2

context into many of today's state-of-the-art natural language processing applications, resulting in generalized checkpoints such as SiEBERT.

As we enter 2023 and beyond, it appears that the primary goals have changed again – current research indicates the focus has moved from classifying information based solely on historical results to using reasoning to prompt new information about the world or determining what's "real" or "fake". Innovations derived from the generative paradigm have yielded unrivaled success across zero-shot scenarios and high-complexity tasks such as detecting sarcasm or fusing multiple modalities together. However, with the advent of these new techniques, substantial challenges associated with Model Variability problems and lack of inter-consistency across models have arisen. The Green AI Initiative has shown the world that high-performance and sustainable systems need not be mutually exclusive by enabling organizations to deploy large-scale models or system-level solutions using 4-bit quantization and targeted distillation.

Current and future research emphasis areas include:

1. Deterministic Benchmarking – establish/testing protocols to measure the consistency of models and ICC in probabilistic settings.

2. Uncertainty-Aware Calibration – integrate Bayesian corrections into MLLM designs for stable functioning within high-risk decision environments such as finance and health care.

Autonomous Multimodal Reasoning: This approach will enhance lightweight MLLM's capability to produce rational reasoning similar to what humans would produce, through multiple different types of data sources. Agentic Explainability: We are working towards developing "neuro-symbolic systems" that result in providing "traceable intelligence" throughout each step of an autonomous decision-making process. This transition in architecture will enable sentiment analysis to be utilized for not only identifying polarity, but also as an important means of understanding how humans communicate in a very complex and rapidly evolving digital and automated world.

REFERENCES

1. M. M. Ali and E. S. Mohamed, "A systematic review of machine learning, deep learning, and natural language processing for sentiment analysis," in *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, Erode, India, 2024, pp. 787-794, doi: 10.1109/ICSSAS64001.2024.10760614.
2. M. A. Alonso, D. Vilares, and C. Gómez-Rodríguez, "Financial sentiment analysis with transformers," *Expert Systems with Applications*, vol. 238, Art. no. 122034, 2024, doi: 10.1016/j.eswa.2023.122034.
3. I. Villanueva-Miranda, Y. Xie, and G. Xiao, "Sentiment analysis in public health: A systematic review of the current state, challenges, and future directions," *Frontiers in Public Health*, vol. 13, Art. no. 1609749, 2025, doi: 10.3389/fpubh.2025.1609749.
4. S. H. Muhammad et al., "SemEval-2025 task 11: Bridging the gap in text-based emotion detection," in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria, 2025. [Online]. Available: https://aclanthology.org/2025.semeval-1.0/
5. S. Vajjala and S. Shimangaud, "Text classification in the LLM era—Where do we stand?," arXiv preprint arXiv:2502.11830, 2025. [Online]. Available: https://arxiv.org/abs/2502.11830
6. J. O. Krugmann and J. Hartmann, "Sentiment analysis in the age of generative AI," *Customer Needs and Solutions*, vol. 11, no. 1, Art. no. 4, 2024, doi: 10.1007/s40547-024-00143-4.
7. E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok, "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," in *HCI International 2024 – Late Breaking Papers*, ser. Lecture Notes in Computer Science, vol. 15382. Cham, Switzerland: Springer, 2024, pp. 197–216, doi: 10.1007/978-3-031-76827-9_11.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, July 2026
Impact Factor: 8.2

8. C. Yan et al., "Sentiment analysis and topic mining using a novel deep attention-based parallel dual-channel model for online course reviews," *Cognitive Computation*, vol. 15, no. 1, pp. 304-322, 2023, doi: 10.1007/s12559-022-10083-7.
9. A. Joshy and S. Sundar, "Analyzing the performance of sentiment analysis using BERT, DistilBERT, and RoBERTa," in *2022 IEEE International Power and Renewable Energy Conference (IPRECON)*, Kollam, India, 2022, pp. 1-6, doi: 10.1109/IPRECON55716.2022.10059542.
10. J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More than a feeling: Accuracy and application of sentiment analysis," International Journal of Research in Marketing, vol. 40, no. 1, pp. 75-87, 2023, doi: 10.1016/j.ijresmar.2022.05.005.
11. D. Ghosh, A. Singh, and R. K. Singh, "Sarcasm detection using deep learning: A comprehensive review," Information Fusion, vol. 91, pp. 312–334, Mar. 2023, doi: 10.1016/j.inffus.2022.10.015. (Replaces original [11])
12. K. O. Adefemi and M. B. Mutanga, "A hybrid deep learning model combining convolutional neural networks (CNN) and long short-term memory (LSTM) networks for sentiment analysis," Digital, vol. 5, no. 2, Art. no. 16, 2025, doi: 10.3390/digital5020016.
13. Siebert, "Sentiment in English RoBERTa-large meta-sentiment model," Hugging Face Hub, 2023. [Online]. Available: https://huggingface.co/siebert/sentiment-roberta-large-english
14. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019. [Online]. Available: https://arxiv.org/abs/1910.01108
15. G. Fatouros, K. Metaxas, J. Soldatos, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with ChatGPT," Machine Learning with Applications, vol. 14, Art. no. 100508, 2023, doi: 10.1016/j.mlwa.2023.100508.
16. PREFEVAL team, "A benchmark for evaluating LLMs' ability to infer and adhere to user preferences," arXiv preprint arXiv:2502.09597, 2025. [Online]. Available: https://arxiv.org/abs/2502.09597
17. Z. Yang et al., "MulCoT-RD: Multimodal sentiment reasoning distillation for JMSRC," arXiv preprint arXiv:2508.05234, 2025. [Online]. Available: https://arxiv.org/abs/2508.05234
18. G. E. Centeno and F. M. Kling, "Irony detection in 19th-century Latin American newspapers utilizing LLMs," arXiv preprint arXiv:2503.22585, 2025. [Online]. Available: https://arxiv.org/abs/2503.22585
19. Y. Zhang, G. Xie, J. Lin, J. Bao, Q. Wang, X. Zeng, and R. Xu, "Targeted distillation for sentiment analysis," in Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), Suzhou, China, Nov. 2025, pp. 22158–22181, doi: 10.18653/v1/2025.emnlp-main.1127. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1127/
20. I. Ahmed, G. S. Khanday, and T. R. Soomro, "Recent applications of explainable AI (XAI): A systematic literature review," Applied Sciences, vol. 14, no. 19, Art. no. 8884, 2024, doi: 10.3390/app14198884.
21. X. Li, G. D. Karampatakis, H. E. Wood, C. J. Griffiths, B. Mihaylova, N. S. Coulson, A. Pasinato, P. Panzarasa, M. Viviani, and A. De Simoni, "The promise of large language models in digital health: Evidence from sentiment analysis in online health communities," arXiv preprint arXiv:2508.14032, 2025. [Online]. Available: https://arxiv.org/abs/2508.14032
22. Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692
23. A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting and analysing financial sentiment," ACM Computing Surveys, vol. 55, no. 13s, Art. no. 274, Jul. 2023, doi: 10.1145/3580485. *(Replaces original [23] with peer-reviewed journal version)*

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, July 2026 Impact Factor: 8.2

24. Radford et al., "Language models are unsupervised multitask learners," OpenAI Technical Report, 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
25. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
26. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
27. T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Oct. 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.
28. L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, Art. no. e1253, 2018, doi: 10.1002/widm.1253.
29. A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," Artificial Intelligence Review, vol. 53, no. 6, pp. 4335–4385, Aug. 2020, doi: 10.1007/s10462-019-09794-5.
30. M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis," Future Generation Computer Systems, vol. 115, pp. 279–294, Feb. 2021, doi: 10.1016/j.future.2020.08.005.
31. X. Zhou, X. Wan, and J. Xiao, "Attention-based LSTM network for cross-lingual sentiment classification," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, Nov. 2016, pp. 247–256, doi: 10.18653/v1/D16-1024.
32. R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, Oct. 2013, pp. 1631–1642. [Online]. Available: https://aclanthology.org/D13-1170/
33. Baziotis, N. Athanasiou, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos, "NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning," in Proceedings of The 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, Jun. 2018, pp. 245–255, doi: 10.18653/v1/S18-1037.
34. H. T. Phan, N. T. Nguyen, and D. Hwang, "Aspect-level sentiment analysis: A survey of graph convolutional network methods," Information Fusion, vol. 91, pp. 149–172, Mar. 2023, doi: 10.1016/j.inffus.2022.10.004.
35. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
36. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, vol. 26, Lake Tahoe, NV, USA, Dec. 2013, pp. 3111–3119.
37. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
38. M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," Computer Science Review, vol. 27, pp. 16–32, Feb. 2018, doi: 10.1016/j.cosrev.2017.10.002.
39. R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," Communications of the ACM, vol. 63, no. 12, pp. 54–63, Dec. 2020, doi: 10.1145/3381831.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, July 2026
Impact Factor: 8.2

40. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al., "HuggingFace's Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Oct. 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.
41. S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in Emotion Measurement, 2nd ed., H. L. Meiselman, Ed. Duxford, UK: Woodhead Publishing, 2021, pp. 323–352, doi: 10.1016/B978-0-12-821124-3.00011-7.
42. A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, Sep. 2017, pp. 1103–1114, doi: 10.18653/v1/D17-1115.
43. 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, Sep. 2017, doi: 10.1016/j.inffus.2017.02.003.
44. L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," arXiv preprint arXiv:1908.03557, 2019. [Online]. Available: https://arxiv.org/abs/1908.03557
45. H. Touvron et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023. [Online]. Available: https://arxiv.org/abs/2302.13971