Advances in Deep Learning for Affective Intelligence: Language Models, Multimodal Trends, and Research Frontiers

Diego Andres Andrade-Segarra[1], Juan Carlos Santillán-Lima[2], Miguel Duque-Vaca[3],
Fernando Tiverio Molina-Granja[4]
Facultad de Informática y Electrónica, Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador[1, 2, 3]
Facultad de Ingeniería, Universidad Nacional de Chimborazo, Riobamba, Ecuador[4]

***Abstract*—The accelerated growth of digital content and the increasing presence of emotional expressions, polarized opinions, and toxic behaviors in social media have driven the development of advanced Affective Analysis techniques. This study presents a broad and up-to-date review of recent studies covering Sentiment Analysis, Emotion Recognition, Hate Speech Detection, cyberbullying, and multimodal approaches grounded in deep learning. The review provides a comparative analysis of the architectures employed—including Transformer-based Models, multimodal frameworks, and variants designed for low-resource languages—along with their metrics, performance outcomes, and emerging patterns. The findings reveal a clear consolidation of Transformer-based Models as the dominant standard, significant progress in multimodality for affective interpretation, and growing attention to multilingual models adapted to diverse cultural contexts. Furthermore, persistent challenges are identified, including limitations related to data availability and quality, Explainable AI (XAI), computational efficiency, and robustness in cross-domain generalization. This review synthesizes current trends, limitations, and opportunities in the field, offering a structured perspective that can serve as a reference for researchers and practitioners involved in the development of more accurate, efficient, and culturally responsible affective systems.**

***Keywords*—Sentiment Analysis; Emotion Recognition; Hate Speech Detection; cyberbullying; deep learning; Transformer-based Models; Multimodal Analysis; multilingual NLP; Explainable AI (XAI); social media**

I. Introduction

Sentiment Analysis, Emotion Recognition, and hate speech classification have become central components in the study of contemporary digital behavior, particularly due to the massive growth of communication on social media platforms [6]. Online interactions generate large volumes of textual content that require advanced tools for their interpretation, leading traditional natural language processing (NLP) approaches to evolve toward deep learning and Transformer-based Models.

The articles reviewed in this study show that early solutions relying on conventional techniques—such as support vector machines, Naïve Bayes, or simple CNNs—have been progressively surpassed by more sophisticated architectures. For example, works such as [3] and [7] illustrate the transition toward deep models capable of more accurately representing contextual dependencies present in the informal language used on social platforms.

The emergence of Transformer-based Models marked a turning point. Studies such as [2], [11], and [19] demonstrate the superiority of Pre-trained Language Models (PLMs), which offer rich and contextualized semantic representations, significantly improving performance in emotional classification, Sentiment Analysis, and Cyberbullying Detection.

Likewise, research focused on specific languages highlights the relevance of contextual adaptation and transfer learning. Works such as [8], [15], and [36] show that linguistic variants of Transformer-based Models are highly effective in languages with moderate or limited resources.

Another key theme identified in the reviewed articles is the advancement of multimodal approaches. Studies such as [30], [31], and [35] reveal the potential of integrating text with images, audio, or contextual signals to improve affective interpretation in more complex and dynamic environments.

Finally, research addressing specific issues such as hate speech and cyberbullying demonstrates strong results through optimized model configurations. Works such as [1] and [16] highlight how combining Transformer-based Models with fine-tuning, feature selection, or hybrid mechanisms substantially improves performance metrics.

Collectively, these advancements show that affective NLP has reached a stage of methodological maturity characterized by the dominance of Transformer architectures, the rise of multimodality, and the adaptation to multilingual contexts. This review synthesizes these developments, offering a structured and up-to-date perspective on the current state of the field and establishing the conceptual foundation for the detailed analysis presented in the subsequent sections.

A. Research Questions

To guide the analysis and synthesis of the reviewed literature, this study addresses the following research questions:

RQ1: What deep learning architectures currently dominate affective intelligence tasks in natural language processing?

RQ2: How have multimodal approaches evolved to enhance emotion and sentiment understanding across different data modalities?

RQ3: What challenges and research gaps persist in multilingual and low-resource affective NLP scenarios?

II. Background and Related Work

Sentiment Analysis, Emotion Recognition, and hate speech identification have established themselves as fundamental areas within Natural Language Processing (NLP), particularly driven by the exponential growth of interactions on social media platforms. The studies reviewed in this work show that the evolution of the field has been characterized by a gradual shift away from traditional lexicon-based and statistical approaches toward the widespread adoption of deep learning architectures and Transformer-based Models. Deep learning has also demonstrated strong predictive capability in other applied domains—such as electricity price forecasting—which has further motivated its adoption for complex sequence and representation learning problems [12].

B. Evolution of Sentiment Analysis and Emotion Recognition

Early approaches based on statistical methods and classical machine learning models were progressively replaced by neural architectures such as LSTM and CNN, which enabled the modeling of sequential dependencies and more complex affective patterns. Studies such as [7] and [3] illustrate this transition, reporting substantial improvements in Affective Analysis tasks through the use of deep neural networks [14].

The introduction of contextual representation techniques, particularly those used in BERT-like models, further strengthened the ability of NLP systems to capture the semantics of informal and noisy social media language. Studies such as [8] and [15] demonstrate the superiority of contextual embeddings over previous methods.

C. Advances Driven by Transformer-Based Models and Pre-Trained Language Models (PLMs)

The consolidation of Transformer-based Models is a recurring pattern across the reviewed literature. Works such as [2], [11], and [19] show how Pre-trained Language Models (PLMs) capture deeper semantic and contextual relationships, resulting in substantial performance gains in emotional classification and Hate Speech Detection. In addition, PLMs such as BERT have been successfully transferred to other domains, including network intrusion detection, underscoring their general-purpose contextual modeling capability [10].

In addition, studies focused on specific languages—such as [36], demonstrate how local adaptations of PLMs can match or even surpass results obtained on larger, more mature corpora.

D. Multimodal Processing and Emerging Research Directions

An emerging trend observable in [29]–[35] is the increasing use of multimodal approaches that integrate text, images, and additional signals. Models such as MemoCMT [30], Cross-modal BERT [31], and Multimodal GRU with Directed Pairwise Cross-Modal Attention [35] show that incorporating multiple modalities can enhance emotional detection and enrich affective understanding in complex environments.

These studies indicate that mechanisms based on Cross-modal Attention, multi-level fusion, and intermodal architectures are expanding NLP toward contexts closer to human perception, such as conversational analysis, emotional video understanding, and monitoring of digital interactions.

E. Multilingualism, Low-Resource Languages, and Knowledge Transfer

The reviewed studies also highlight the importance of multilingualism and the challenges associated with low-resource languages. Works such as [16] and [13] emphasize the need to adapt models and transfer learning strategies to specific linguistic contexts.

Moreover, the evidence shows that models such as IndoBERT, RoBERTa, and domain-adapted variants can compensate for the lack of large annotated corpora, achieving competitive results when properly optimized and contextually fine-tuned.

F. Synthesis of the State-of-the-Art

Based on the analysis of the reviewed studies, the following dominant trends are identified:

- A clear predominance of Transformer-based Models as the primary architecture.
- A sustained growth of multimodal approaches for Affective Analysis.
- Extensive use of fine-tuning techniques, knowledge transfer, and linguistic variants of Pre-trained Language Models (PLMs).
- Consolidation of hybrid strategies integrating deep neural networks, graphical representations, and explainable mechanisms. (e.g., transformer–GNN convergence frameworks [9]).
- Increasing attention to low-resource languages and specialized domains through optimization and Domain Adaptation.

This conceptual framework synthesizes the current state of affective NLP, according to the evidence provided by the analyzed studies, and serves as the foundation for the discussion, comparisons, and conclusions presented in the following sections.

III. Methodology

A. Methodological Positioning

This work is positioned as a narrative, corpus-driven review rather than a fully systematic review as defined by formal PRISMA-style protocols. While a structured Scopus-based search strategy was employed to identify relevant literature, the final corpus was refined through controlled inclusion criteria focused on thematic relevance, publication impact, and methodological diversity. The objective of this review is not exhaustive coverage, but rather a comprehensive and interpretative synthesis of recent advances in deep learning–based affective intelligence, with particular emphasis on language models, multimodal architectures, and emerging research trends.

B. General Review Approach

The methodology employed in this review was specifically designed to analyze, compare, and synthesize the contributions of the selected studies, all focused on Sentiment Analysis, Emotion Recognition, hate speech classification, and multimodal or multilingual approaches within the context of Natural Language Processing (NLP). Consistent with the methodological positioning described above, the present study adopts an empirical, corpus-driven approach grounded exclusively in the explicit, verifiable characteristics of the included studies.

The methodological process was structured into three specific stages derived exclusively from information present in the studies:

1) Identification and compilation of the corpus: Each study was selected for addressing one or more of the target themes—sentiment, emotion, hate speech, multimodality, or multilingualism—and for being part of the predefined document base.

2) Characterization and systematic extraction of information: Each publication was examined in detail to extract consistent data regarding architectures, datasets, preprocessing techniques, evaluation metrics, languages, domains, and experimental strategies.

3) Comparative and thematic synthesis: Articles were grouped according to observed patterns—text-based models, multimodal approaches, hate speech or cyberbullying detection, multilingual methods, and Explainable AI (XAI)—to generate an integrated interpretation of the current state of the field.

This approach ensures that all methodological elements are grounded exclusively in the analyzed corpus, without relying on external methodological frameworks unrelated to the content of the articles.

C. Selection and Composition of the Corpus

The corpus used in this review comprises studies published between 2018 and the present, ensuring a representative coverage of recent developments in Sentiment Analysis, Emotion Recognition, Hate Speech Detection, and multimodal deep learning approaches.

Although the set of studies was consolidated manually for this review, the original document pool stems from three major academic sources used during the initial collection process:

- Scopus (Elsevier): multidisciplinary database for identifying recent and relevant studies.
- IEEE Xplore: a primary source for research in deep learning, NLP, and Affective Analysis.
- Elsevier ScienceDirect: repository for applied studies related to emotion, sentiment, and hate speech.

A representative Scopus search query was:

(TITLE-ABS-KEY ("sentiment analysis" OR "opinion mining" OR "emotion recognition"

OR "emotion classification" OR "hate speech detection"

OR "cyberbullying detection")

AND

TITLE-ABS-KEY ("deep learning" OR "neural network*" OR "transformer*"

OR "BERT" OR "RoBERTa" OR "GPT" OR "multimodal"

OR "cross-modal")

)

AND (PUBYEAR > 2017 AND PUBYEAR < 2026)

AND (LIMIT-TO (DOCTYPE , "ar"))

AND (LIMIT-TO (LANGUAGE , "English") OR LIMIT-TO (LANGUAGE , "Spanish"))

These initial searches—along with complementary retrieval in IEEE Xplore and ScienceDirect—yielded a broad collection of publications that were manually filtered to form the final corpus.

The final set of studies did not arise from a purely database-driven systematic search, but from a deliberate, controlled selection in which each article was included based on the following explicit and observable criteria:

- Application or evaluation of Deep Learning Approaches to affective tasks.
- Focus on Sentiment Analysis, Emotion Recognition, Hate Speech Detection, cyberbullying, Multimodal Learning, or multilingual models.
- Reporting quantitative experimental results (accuracy, precision, recall, F1-score, etc.).
- Clear description of architecture, components, and training strategies.
- Verifiable information on datasets, languages, and domains.

Articles lacking empirical evidence, purely theoretical works, or non-peer-reviewed materials were excluded.

These adaptations constitute the structural basis of the instrument described in Section C.

D. Data Extraction and Structuring

Data extraction was conducted manually and organized according to common elements identified across all the studies:

- Architectures: Transformer-based Models (BERT, RoBERTa, IndoBERT), CNN, LSTM, BiGRU, GNN, hybrid systems, and Multimodal Learning.
- Preprocessing: tokenization, normalization, noise removal, static and contextual embeddings.
- Datasets: corpora in English, Spanish, Indonesian, and others; social media datasets (Twitter, Facebook, Google Play); hate-speech, emotion, and multimodal collections.
- Training strategies: fine-tuning, hyperparameter optimization, lightweight models, multimodal fusion.

- Metrics: F1-score, accuracy, precision, recall, macro-F1, weighted-F1.
- Key contributions: performance gains, new architectures, fusion techniques, language-specific enhancements, use of XAI.
- Limitations: dataset bias, class imbalance, computational demands, small corpora, and lack of Explainable AI (XAI).

Each dimension was documented consistently to enable a balanced, systematic comparison.

E. Thematic Classification Derived from the Studies

The categories in Table I emerge naturally from methodological, architectural, and experimental patterns identified across the corpus and illustrate how the current landscape of affective NLP is organized and how each research line contributes to the field.

TABLE I. THEMATIC CLASSIFICATION OF THE REVIEWED STUDIES

Thematic Category	Representative Articles	Key Findings
Text-based Sentiment Analysis	[2], [3], [7], [8], [11], [15]	Transformers outperform LSTM/CNN in informal language
Emotion and Affective State	[13], [15], [30]–[35]	Multimodality increases emotional detection accuracy
Hate Speech and Cyberbullying	[1], [16], [18], [19], [20], [21]	Optimized and lightweight models achieve strong results
Multilingual / Low-resource Models	[13], [16], [19], [36]	Transfer learning improves performance in low-resource settings
Explainable AI (XAI) Models	[16], [21], [27]	Interpretability is increasingly important in critical applications

Source: Author's elaboration.

F. Ethical Considerations and Transparency

All included articles were reviewed to ensure that their procedures were described with enough clarity to allow understanding and reproducibility. This review maintains methodological transparency by relying exclusively on explicit information contained within all the studies, without incorporating external methodological frameworks.

IV. RESULTS AND DISCUSSION

This section synthesizes and comparatively analyzes the findings reported across all the studies, considering the architectures employed, performance metrics, linguistic domains, multimodal approaches, and specific applications such as Hate Speech Detection and cyberbullying. The analysis is structured according to patterns observed directly in the studies.

A. Performance of Text-Based Models

Text-based models constitute the core of affective NLP across the reviewed studies. The results consistently show that Transformer-based Models outperform classical deep learning architectures in tasks such as Sentiment Analysis, Emotion Recognition, and text-based Hate Speech Detection. In particular, BERT and its variants achieve superior accuracy and F1-score due to their ability to capture contextual semantic relationships, as evidenced in [2], [7], [8], [11], [15], [16], [19], and [36]. Classical architectures, such as LSTM, CNN, or BiGRU, used in studies like [3], [7], [11], and [15], provide competitive baselines and perform adequately in small or well-structured datasets, but their capacity to model complex informal language is more limited. Hybrid text-based models—such as BERT combined with BiGRU or GNN components in [11]—demonstrate improvements in stability and generalization, indicating that architectural fusion can yield benefits even when working exclusively with textual input. Overall, the evidence confirms that contextualized Transformer-based Models represent the dominant paradigm in text-only affective tasks, while traditional and hybrid architectures remain useful in scenarios with constrained computational resources or smaller datasets. Aspect-based sentiment analysis on microblogging and review platforms also benefits from contextual Transformer representations, which better capture fine-grained opinion targets and their polarity [22]. Similarly, CNN-based pipelines enriched with RoBERTa embeddings have shown competitive performance in sentiment classification for movie reviews and other noisy user-generated text [4]. For Emotion Recognition, RoBERTa variants enhanced with emotion-aware attention mechanisms report additional gains by focusing representation learning on affective cues [23].

B. Models for Hate Speech Detection and Cyberbullying

Articles [1], [16], [18], [19], [20], [21] focus on detecting toxic speech, hate speech, or cyberbullying.

Their findings consistently show that BERT-based models [1], [16], [19] significantly outperform classical architectures. In contrast, [18] combines GloVe + PCA with a Transformer, achieving efficient fusion of static and contextual embeddings. Alternatively, [20] proposes a lightweight architecture (DBFN-J) with strong performance and lower computational cost and [21] provides a comprehensive review of LLMs for hate speech, highlighting multilingual robustness.

Overall, these works indicate that the most robust models are those balancing performance and efficiency—particularly relevant in sensitive social contexts.

C. Multilingual Models and Applications in Low-Resource Languages

Studies addressing Sentiment Analysis, Emotion Recognition, and text classification across different languages reveal a consistent trend in which Transformer-based Models outperform LSTM, CNN, and hybrid architectures.

For example, domain-specific fine-tuning of BERT in [2] leads to clear improvements in accuracy and F1-score, while [3] demonstrates that Deep Learning approaches surpass classical techniques when analyzing large-scale Twitter data. Similarly, [7], [8], and [15] show that BERT variants adapted to specific domains and languages more effectively capture informal and context-dependent semantics, and [11] integrates BERT with BiGRU and GCN to achieve additional gains in generalization. Collectively, these studies indicate that contextualized text-based models consistently deliver the strongest affective performance. Related multilingual emotion extraction studies further confirm the importance of language-specific resources and cross-lingual modeling choices [17].

At the multilingual level, research highlights the particular challenges posed by low-resource settings. In [13], the authors analyze Emotion Recognition through a cross-cultural perspective, showing that sociocultural factors influence affective interpretation. In [16], the authors demonstrate that fine-tuned Transformer-based Models can overcome data scarcity barriers, while [19] evaluates multilingual LLMs for Hate Speech Detection across diverse linguistic contexts. Furthermore, [36] adapts IndoBERT to Indonesian emotional content, reaching performance levels comparable to models trained on larger or better-resourced corpora.

Overall, multilingual fine-tuning, transfer learning, and linguistic adaptation emerge as essential strategies for extending the applicability of affective NLP to culturally diverse and resource-constrained environments.

D. Multimodal Learning

Articles [29], [30], [31], [32], [33], [34], [35] present multimodal architectures integrating text with images, audio, or other signals.

Common trends include Cross-modal Attention and multi-level fusion — [29], [33], Multimodal Transformer-based models — [30], [31], Hierarchical and intramodal enhancement — [32], [34], and GRU networks with directed cross-modal attention — [35].

Results consistently show that multimodal approaches outperform text-only systems when emotional cues depend on visual, acoustic, or combined signals.

E. Global Comparison of Architectures and Metrics

Table II shows how the main architectures used across the studies are distributed, highlighting the predominance of Transformer-based Models and the emerging role of Multimodal Learning.

TABLE II. COMPARATIVE SUMMARY OF ARCHITECTURES USED ACROSS THE STUDIES

Model Category	Associated Articles	Key Trends	Common Metrics
Transformers (BERT, RoBERTa, IndoBERT, LLMs)	[2], [7], [8], [11], [15], [16], [19], [36]	Best overall performance in affective and social tasks	F1-score, accuracy
Classical DL Models (LSTM, CNN, BiGRU)	[3], [7], [11], [15]	Lower performance vs. Transformers; useful in small datasets	accuracy, F1-score
Hybrid Models (BERT+GRU, PCA+GloVe, GNN)	[11], [18], [20]	Balanced trade-off between efficiency and precision	macro-F1, weighted-F1
Multimodal Models	[29]–[35]	Improved affective understanding when combining signals	F1-score, weighted-F1

Source: Author's elaboration.

A global comparison of the architectures and evaluation metrics used across all the studies reveals several consistent patterns. Transformer-based Models dominate performance in all affective tasks, reflecting their superior ability to capture contextual dependencies. Multimodal Learning further enhances performance in enriched contexts where affective cues extend beyond text. Lightweight Models such as the one proposed in [20], demonstrate strong competitive performance when computational efficiency is a priority. Across the corpus, the most commonly reported metrics include F1-score, accuracy, and macro-F1, underscoring their relevance in affective evaluation. Finally, dataset quality—in terms of size, balance, and language—proves to be as influential as the choice of architecture, confirming its central role in determining model robustness and generalization.

F. Synthesis of Global Patterns Across the Corpus

Across the reviewed studies:

- Transformers establish a new state-of-the-art in affective NLP.
- Multimodality represents the next major step in emotional interpretation.
- Multilingualism remains challenging, but is increasingly well addressed through adapted PLMs.
- For hate speech and cyberbullying, hybrid and lightweight models provide excellent precision–efficiency trade-offs.

To provide a concise and integrative overview of the main findings discussed in this section, Table III summarizes the relationships between affective tasks, representative models, datasets, and key outcomes identified across the reviewed studies.

TABLE III. SYNTHETIC OVERVIEW OF TASKS, MODELS, DATASETS, AND KEY FINDINGS

Task	Representative Models	Typical Datasets	Key Findings
Sentiment Analysis	BERT, RoBERTa, CNN-LSTM	Twitter, Reviews	Transformer-based models consistently outperform classical deep learning approaches
Emotion Recognition	BERT, Multimodal Transformers	EmoBank, IEMOCAP	Multimodal fusion improves affective accuracy and robustness
Hate Speech Detection	BERT, DBFN-J	Twitter, HASOC	Fine-tuned pre-trained language models balance performance and computational efficiency
Multilingual Affective NLP	IndoBERT, mBERT	Multilingual corpora	Transfer learning mitigates data scarcity in low-resource settings

Source: Author's elaboration.

V. CURRENT CHALLENGES AND FUTURE RESEARCH DIRECTIONS

A. Current Challenges Observed in the Studies

Table IV synthesizes the main challenges explicitly documented in the corpus, allowing the identification of cross-cutting patterns that affect the development of affective NLP

systems in real-world contexts. Beyond data limitations, linguistic phenomena such as sarcasm and irony remain difficult to model reliably and continue to impact affective classification performance in social media contexts [5]. Moreover, recent data-centric and human-in-the-loop error analysis frameworks supported by Explainable AI provide practical strategies to diagnose and mitigate systematic model failures [28].

TABLE IV. CURRENT CHALLENGES IDENTIFIED ACROSS THE STUDIES

Challenge	Articles Reporting It	Observations
Corpus scarcity/limitations	[3], [7], [8], [13], [15], [16], [29]–[35], [36]	Issues related to size, imbalance, and domain specificity
Linguistic complexity	[1], [3], [7], [8], [13], [15], [36]	Difficulty handling sarcasm, social variation, cultural differences
High computational costs	[20], [29]–[35]	Multimodal and heavy Transformer-based models
Lack of explainability	[20], [29]–[35]	Critical issue in sensitive environments
Inconsistent evaluation	[2], [11], [18], [29]	Non-comparable metrics; use of different datasets

Source: Author's elaboration.

B. Limitations Identified in the Studies

The consolidated analysis reveals a set of recurrent limitations affecting the development, evaluation, and generalization capacity of Affective Analysis systems applied to social media. These limitations fall into five main categories. Inconsistent evaluation protocols and limited cross-dataset testing hinder comparability across studies; similar concerns have motivated unified benchmark construction in other AI areas [25] and cross-dataset analyses of language models to assess robustness and distributional shift [26].

In terms of data quality and availability, several studies highlight structural weaknesses in the datasets. Articles [3], [7], [8], [13], [15], and [36] report datasets that are small, imbalanced, or highly domain-specific, limiting model generalization. In contrast, article [16] emphasizes that Low-resource Languages suffer from limited annotation, hindering both training and validation. Meanwhile, Multimodal studies [29]–[35] report the cost and complexity of obtaining synchronized text–image–audio datasets, which restricts both scope and reproducibility. Class imbalance is particularly recurrent in toxic speech and emotion datasets, where oversampling and related resampling strategies are often required to stabilize training and improve minority-class performance [24].

VI. CONTRIBUTIONS TO THE FIELD

This review advances the state of knowledge in affective intelligence by providing a broad and up-to-date integrated examination of deep learning, multimodal fusion, and large-scale language models through a unified analytical framework. While prior studies typically address sentiment analysis, emotion recognition, or multimodal architectures in isolation, this study synthesizes cross-domain evidence to identify convergent methodological patterns, emerging architectures, and persistent research gaps across text, audio, and visual modalities. The work contributes novel comparative insights into how transformer-based, hybrid, and cross-modal embedding strategies are reshaping affective computing, and outlines a forward-looking taxonomy of trends that can inform the design of next-generation AI systems capable of nuanced emotional reasoning. By bridging findings from NLP, computer vision, and affective signal processing, this review positions itself as a foundational reference for researchers developing emotionally aware AI technologies.

VII. GENERAL CONCLUSIONS

This corpus-driven review of contemporary studies in Affective Analysis—including Sentiment Analysis, Emotion Recognition, Hate Speech Detection, cyberbullying, and multimodal approaches—allows us to establish solid conclusions regarding the current state of the field, its dominant trends, and the impact of deep learning–based models. This synthesis draws exclusively from the findings discussed in the previous sections and remains fully aligned with the methodology adopted throughout the study.

A. Transformation of Affective NLP Through Transformer-Based Models

Text-based models constitute the core of affective NLP across the reviewed studies. The results consistently show that Transformer-based Models outperform classical deep learning architectures in tasks such as Sentiment Analysis, Emotion Recognition, and text-based Hate Speech Detection. In particular, BERT and its variants achieve superior accuracy and F1-score due to their ability to capture contextual semantic relationships, as evidenced in [2], [7], [8], [11], [15], [16], [19], and [36]. Classical architectures such as LSTM, CNN, or BiGRU, used in studies like [3], [7], [11], and [15], provide competitive baselines and perform adequately in small or well-structured datasets, but their capacity to model complex informal language is more limited. Hybrid text-based models—such as BERT combined with BiGRU or GNN components in [11]—demonstrate improvements in stability and generalization, indicating that architectural fusion can yield benefits even when working exclusively with textual input. Overall, the evidence confirms that contextualized Transformer-based Models represent the dominant paradigm in text-only affective tasks, while traditional and hybrid architectures remain useful in scenarios with constrained computational resources or smaller datasets.

B. Multimodality as a High-Impact Emerging Axis

The multimodal works — [29]–[35] — show that integrating text with images, audio, or additional signals significantly enhances the ability to interpret emotions, especially in contexts where textual information alone is ambiguous or insufficient. These studies consolidate multimodality as a key trend in the evolution of Affective Analysis.

C. Relevance of Multilingualism and Cultural Adaptation

Studies targeting languages other than English — [13], [15], [16], [19], [36]— highlight the importance of considering sociocultural factors and data availability. The results show that multilingual or domain-adapted models can achieve competitive

performance when supported by representative datasets and appropriate fine-tuning strategies.

D. Need for Improved Data, Explainable AI (XAI), and Efficiency

The integrated analysis of the corpus reveals three persistent cross-cutting needs:

Improved and expanded datasets

Reported in [3], [7], [8], [13], [15], [16], [36], the limitations in dataset size, balance, and diversity continue to affect generalization.

Explainable AI (XAI)

Articles [16], [19], [21], [27] emphasize the need to incorporate XAI mechanisms to ensure trust and transparency.

Computational optimization

Studies such as [20] and multimodal works [29]–[35] underline the importance of developing lighter models with lower computational requirements to enable real-world deployment.

E. Central Contribution of this Review

This review provides a structured, comparative, and up-to-date perspective on affective NLP, highlighting how recent advancements converge toward:

highly contextualized architectures,

integration of multiple information modalities,

models adaptable to diverse languages and cultures,

growing interest in transparency and efficiency.

This synthesis consolidates a deep understanding of the dominant trends and remaining challenges in the field, serving as a valuable reference for researchers and practitioners developing robust and reliable affective systems.

The findings of this review underscore a clear evolution toward emotion-aware AI models that integrate deep learning, multimodal fusion, and self-supervised language architectures. Although current systems demonstrate impressive performance across benchmark datasets, their real-world applicability remains limited by challenges such as contextual ambiguity, cultural variation in emotional expression, dataset bias, and the difficulty of modeling complex affective states beyond basic emotions.

F. Future Research Directions

Future research should advance in at least three directions.

First, there is a need for large-scale, culturally diverse multimodal datasets that reflect real-world emotional variability and enable robust generalization across languages, accents, and socio-cultural contexts.

Second, the emergence of multimodal LLMs calls for new architectures capable not only of classifying emotions but also of explaining the reasoning behind affective predictions, thus integrating affective computing with explainable AI.

Third, next-generation models should incorporate continuous affective dynamics—such as stress, fatigue, ambiguity, and mixed emotions—moving beyond categorical labels toward richer temporal representations aligned with human psychology.

Collectively, these research directions provide a roadmap for building emotionally intelligent AI systems that are reliable, transparent, and more deeply aligned with human communication.

Declaration on Generative AI

The authors used generative Artificial Intelligence tools solely to support textual coherence revision, improve writing quality, and optimize English translations during the preparation of this manuscript. Their use was strictly limited to linguistic assistance and did not intervene in the conceptual development, analysis, interpretation of results, selection or evaluation of the corpus, or the elaboration of the scientific contributions.

References

[1] M. Mubeen, A. Muskan, A. Akram, J. Rashid, T. A. N. Alshalali, and N. Sarwar, "Cyberbullying-Related Automated Hate Speech Detection on Social Media Platforms Using Stack Ensemble Classification Method," International Journal of Computational Intelligence Systems, vol. 18, no. 1, p. 174, 2025, doi: 10.1007/s44196-025-00919-z.

[2] U. K. Das, R. S. Ani, N. Datta, I. Fahad, J. Sikder, U. Sara, and A. Chakraborty, "Enhancing Sentiment Analysis Accuracy on Social Media Comments Using a Tuned BERT Model," Discover Computing, vol. 28, no. 1, p. 198, 2025, doi: 10.1007/s10791-025-09599-x.

[3] S. Darad and S. Krishnan, "Sentimental Analysis of COVID-19 Twitter Data Using Deep Learning and Machine Learning Models," Ingenius. Revista de Ciencia y Tecnología, no. 29, pp. 108–117, 2023, doi: 10.17163/ings.n29.2023.10.

[4] B. Paneru, B. Thapa, and B. Paneru, "Sentiment Analysis of Movie Reviews: A Flask Application Using CNN With RoBERTa Embeddings," Systems and Soft Computing, vol. 7, p. 200192, 2025, doi: 10.1016/j.sasc.2025.200192.

[5] M. Madhavi, C. R. M. Reddy, P. K. Mannepalli, S. S., V. Sravanthi, L. P. Maguluri, and U. G. Naidu, "Adaptive Bi-Directional Long Short-Term Memory-Based Sarcasm Detection on Social Media Platforms," Discover Computing, vol. 28, no. 1, p. 214, 2025, doi: 10.1007/s10791-025-09722-y.

[6] F. A. Lovera and Y. Cardinale, "Sentiment Analysis in Twitter: A Comparative Study," Revista Científica de Sistemas e Informática, vol. 3, no. 1, p. e418, 2023, doi: 10.51252/rcsi.v3i1.418.

[7] D. Andrade-Segarra and G. León-Paredes, "Deep Learning-Based Natural Language Processing Methods Comparison for Presumptive Detection of Cyberbullying in Social Networks," International Journal of Advanced Computer Science and Applications, vol. 12, no. 5, pp. 259–268, 2021, doi: 10.14569/IJACSA.2021.0120592.

[8] J. J. López Condori and F. O. Gonzales Saji, "Análisis de Sentimiento de Comentarios en Español en Google Play Store Usando BERT," Ingeniare. Revista chilena de ingeniería, vol. 29, no. 4, pp. 557–563, 2021.

[9] S. Chu and J. Liu, "Based on BERT–GPT–GNN Converged Architecture: Intelligent Generation Engine for Complex SQL Queries in Business Intelligence," Discover Artificial Intelligence, vol. 5, no. 1, p. 147, 2025, doi: 10.1007/s44163-025-00381-y.

[10] Y. Yang and X. Peng, "BERT-Based Network for Intrusion Detection System," EURASIP Journal on Information Security, vol. 2025, no. 1, p. 11, 2025, doi: 10.1186/s13635-025-00191-w.

[11] M. R. R. Rana, A. Nawaz, S. U. Rehman, M. A. Abid, M. Garayevi, and J. Kajanová, "BERT-BiGRU-Senti-GCN: An Advanced NLP Framework for Analyzing Customer Sentiments in E-Commerce," International Journal of Computational Intelligence Systems, vol. 18, no. 1, p. 21, 2025, doi: 10.1007/s44196-025-00747-1.

[12] J. M. Failing, J. Segarra-Tamarit, J. Cardo-Miota, and H. Beltran, "Deep Learning-Based Prediction Models for Spot Electricity Market Prices in the Spanish Market," Mathematics and Computers in Simulation, vol. 240, pp. 96–104, 2026, doi: 10.1016/j.matcom.2025.07.010.

[13] L. Liang and S. Wang, "Spanish Emotion Recognition Method Based on Cross-Cultural Perspective," Frontiers in Psychology, vol. 13, p. 849083, 2022, doi: 10.3389/fpsyg.2022.849083.

[14] M. T. S. Al-Baity and M. D. Rajab, "Computational Linguistics-Based Emotion Detection and Classification Model on Social Networking Data," Applied Sciences, vol. 12, no. 19, p. 9680, 2022, doi: 10.3390/app12199680.

[15] F. M. Plaza-Del-Arco, M. Martín-Valdivia, L. López, and R. Mitkov, "Improved Emotion Recognition in Spanish Social Media Through Incorporation of Lexical Knowledge," Future Generation Computer Systems, vol. 110, pp. 1000–1010, 2020, doi: 10.1016/j.future.2019.09.034.

[16] E. Fetahi, A. Susuri, M. Hamiti, Z. Kastrati, E. Canhasi, and A. Misini, "Enhancing Social Media Hate Speech Detection in Low-Resource Languages Using Transformers and Explainable AI," Social Network Analysis and Mining, vol. 15, no. 1, p. 82, 2025, doi: 10.1007/s13278-025-01497-w.

[17] V. K. Jain, S. Kumar, and S. L. Fernandes, "Extraction of Emotions From Multilingual Text Using Intelligent Text Processing and Computational Linguistics," Journal of Computational Science, vol. 21, pp. 316–326, 2017, doi: 10.1016/j.jocs.2017.01.010.

[18] M. Umer, E. A. Alabdulqader, A. A. Alarfaj, L. Cascone, and M. Nappi, "Cyberbullying Detection Using PCA Extracted GLOVE Features and RoBERTaNet Transformer Learning Model," IEEE Transactions on Computational Social Systems, vol. 12, no. 5, pp. 3881–3890, 2025, doi: 10.1109/TCSS.2024.3422185.

[19] M. Ahmad, M. Waqas, A. Hamza, S. Usman, I. Batyrshin, and G. Sidorov, "UA-HSD-2025: A Large Language Model-Based Approach for Multilingual Hate Speech Detection From Tweets Using Pre-Trained Transformers," Computers, vol. 14, no. 6, p. 239, 2025, doi: 10.3390/computers14060239.

[20] N. F. Janbi, A. A. Almazroi, and N. Ayub, "DBFN-J: A Lightweight and Efficient Model for Hate Speech Detection on Social Media Platforms," International Journal of Advanced Computer Science and Applications, vol. 16, no. 1, 2025, doi: 10.14569/IJACSA.2025.01601128.

[21] A. Albladi et al., "Hate Speech Detection Using Large Language Models: A Comprehensive Review," IEEE Access, vol. 13, pp. 20871–20892, 2025, doi: 10.1109/ACCESS.2025.3532397.

[22] D. Drašković and S. Milanović, "Aspect-Based Sentiment Analysis of User-Generated Content From a Microblogging Platform," Journal of Big Data, vol. 12, no. 1, p. 186, 2025, doi: 10.1186/s40537-025-01244-0.

[23] F. Alqarni, A. Sagheer, A. Alabbad, and H. Hamdoun, "Emotion-Aware RoBERTa Enhanced With Emotion-Specific Attention and TF-IDF Gating for Fine-Grained Emotion Recognition," Scientific Reports, vol. 15, no. 1, p. 17617, 2025, doi: 10.1038/s41598-025-99515-6.

[24] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A Comprehensive Evaluation of Oversampling Techniques for Enhancing Text Classification Performance," Scientific Reports, vol. 15, no. 1, p. 21631, 2025, doi: 10.1038/s41598-025-05791-7.

[25] J. Wang, S. Zou, Y. Wang, W. Huang, and J. Song, "Unified Benchmark Construction and Algorithm Performance Evaluation for Large-Scale Object Detection," Discover Computing, vol. 28, no. 1, p. 196, 2025, doi: 10.1007/s10791-025-09707-x.

[26] D. Garcés, M. Santos, and D. Fernández-Llorca, "Cross-Dataset Analysis of Language Models for Generalised Multi-Label Review Note Distribution in Animated Productions," International Journal of Computational Intelligence Systems, vol. 18, no. 1, p. 88, 2025, doi: 10.1007/s44196-025-00785-9.

[27] M. Norval and Z. Wang, "Explainable Artificial Intelligence Techniques for Speech Emotion Recognition: A Focus on XAI Models," Inteligencia Artificial, vol. 28, pp. 85–123, 2025, doi: 10.4114/intartif.vol28iss76pp85-123.

[28] A. El-Sayed et al., "A Data-Centric HitL Framework for Conducting a Systematic Error Analysis of NLP Datasets Using Explainable AI," Scientific Reports, vol. 15, no. 1, p. 30406, 2025, doi: 10.1038/s41598-025-13452-y.

[29] B. Miao and C. Xu, "Aspect-Level Multimodal Sentiment Analysis Model Based on Multi-Scale Feature Extraction," Scientific Reports, vol. 15, no. 1, p. 31591, 2025, doi: 10.1038/s41598-025-16051-z.

[30] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "MemoCMT: Multimodal Emotion Recognition Using Cross-Modal Transformer-Based Feature Fusion," Scientific Reports, vol. 15, no. 1, p. 5473, 2025, doi: 10.1038/s41598-025-89202-x.

[31] J. Feng, "Cross-Modal BERT Model for Enhanced Multimodal Sentiment Analysis in Psychological Social Networks," BMC Psychology, vol. 13, no. 1, p. 1081, 2025, doi: 10.1186/s40359-025-03443-z.

[32] Y. Cai, X. Li, Y. Zhang, J. Li, F. Zhu, and L. Rao, "Multimodal Sentiment Analysis Based on Multi-Layer Feature Fusion and Multi-Task Learning," Scientific Reports, vol. 15, no. 1, p. 2126, 2025, doi: 10.1038/s41598-025-85859-6.

[33] S. Zhao, J. Ren, and X. Zhou, "Cross-Modal Gated Feature Enhancement for Multimodal Emotion Recognition in Conversations," Scientific Reports, vol. 15, no. 1, p. 30004, 2025, doi: 10.1038/s41598-025-11989-6.

[34] Z. Zhang, Y. Wang, J. Cui, and H. Zheng, "A Multimodal Emotion Recognition Model With Intra-Modal Enhancement and Inter-Modal Interaction," IEICE Transactions on Information and Systems, vol. E108.D, 2025, doi: 10.1587/transinf.2024EDP7161.

[35] Z. Qin, Q. Luo, Z. Zang, and H. Fu, "Multimodal GRU With Directed Pairwise Cross-Modal Attention for Sentiment Analysis," Scientific Reports, vol. 15, no. 1, p. 10112, 2025, doi: 10.1038/s41598-025-93023-3.

[36] C. Shaw, P. LaCasse, and L. Champagne, "Exploring Emotion Classification of Indonesian Tweets Using Large Scale Transfer Learning via IndoBERT," Social Network Analysis and Mining, vol. 15, no. 1, p. 22, 2025, doi: 10.1007/s13278-025-01439-6.