When Text Is Not Enough: Structural Limits of Text-Only Transformer-Based Emotion Classification

Szymon Chirowski,[1] Maciej Czerniak,[1]
Ondrej Mitas,[2] Maks Burchard[1]

[1] Breda University of Applied Sciences, Applied Data Science & AI, Breda, the Netherlands
242621@buas.nl, 243552@buas.nl, 240894@buas.nl
[2] Breda University of Applied Sciences, Tourism, Breda, the Netherlands
mitas.o@buas.nl

This study investigates whether limitations observed in text-only transformer-based emotion classification pipelines reflect implementation shortcomings or structural constraints inherent to unimodal modeling. A pipeline was constructed using unscripted dialogue from *MasterChef Polska*, incorporating automated speech-to-text transcription, neural machine translation, and benchmarking across SVM, Bi-LSTM, and RoBERTa architectures. While the fine-tuned RoBERTa model achieved substantially higher accuracy (0.755), confusion matrix analysis and explainable AI techniques revealed persistent structural asymmetries, including uneven performance across emotion categories, high-arousal anger-joy confusion, and translation-induced distortions. Evaluation against automated labels further exposed a "Ground Truth Paradox," where models are validating each other rather than a human-verified set of conclusions. Increased architectural capacity improves performance but does not resolve structural limitations of text-only emotion classification.

DOI
https://doi.org/10.18690/um.fov.4.2026.44

ISBN
978-961-299-152-4

Keywords:
emotion recognition, affective computing, transformer models, multimodal limitations, explainable AI

University of Maribor Press

1 Introduction

The growing integration of Artificial Intelligence systems into digital media environments has intensified interest in automated emotion recognition. Across domains such as marketing analytics, media monitoring, and audience engagement measurement, quantifying affect at scale is often presented as a cost-efficient alternative to manual interpretation. Text-based Natural Language Processing pipelines, particularly those built on transformer architectures such as Robustly Optimized BERT Approach (RoBERTa), are widely adopted due to their scalability and relatively low operational overhead (Schuller et al., 2026; Wu et al., 2025). However, reliance on unimodal, text-only systems raises fundamental questions about theoretical validity. Emotional meaning in real-world interactions rarely emerges from language alone; affective interpretation depends on vocal prosody, facial expression, interpersonal dynamics, and situational context (Picard, 2000; Wu et al., 2025). When classification pipelines rely exclusively on textual input, these multimodal cues are necessarily excluded.

The technical design of emotion recognition systems reflects underlying psychological assumptions. Many contemporary pipelines adopt discrete categorical taxonomies derived from Basic Emotion Theory, proposing universal categories such as joy, sadness, anger, fear, surprise, and disgust (Ekman, 1992; Schuller et al., 2026). These predefined classes align conveniently with supervised machine learning, where labelled examples are mapped to fixed outputs (Azari et al., 2020). From this perspective, emotions are treated as relatively stable patterns inferable from linguistic markers.

In contrast, constructionist accounts conceptualize affect as context-dependent and dynamically assembled rather than biologically fixed (Barrett, 2017). Emotional experience emerges through the integration of language, bodily regulation, prior knowledge, and environmental context. If affect is constructed through multimodal interaction, restricting analysis to textual signals may impose structural constraints on classification accuracy (Wu et al., 2025). This tension between categorical convenience and contextual complexity has direct implications for the reliability of text-only emotion recognition systems.

Rather than treating misclassification as an implementation defect, the analysis examines whether certain limitations stem from structural properties inherent to unimodal text modelling. In doing so, the paper contributes to ongoing discussions about the validity, reliability, and governance of automated affect recognition systems. As regulatory frameworks such as the EU AI Act (Regulation (EU) 2024/1689, 2024) increasingly scrutinize high-risk AI applications, understanding the boundaries of text-only emotion classification becomes both a technical and ethical concern. If incorrect AI-based coding of affect is a substantial risk, it is important to identify these errors and their boundary conditions.

2 Theoretical Background: The Conflict of Affective Frameworks

Automated emotion recognition systems operationalize specific assumptions about the nature of affect. In most computational implementations, emotion classification is framed as a supervised learning task in which observable signals are mapped to predefined categorical labels (Azari et al., 2020; Schuller et al., 2026). This structure reflects the basic emotion framework, which conceptualizes emotions as discrete, biologically grounded states distinguishable through identifiable expressive patterns (Ekman, 1992). Within this paradigm, categories such as anger, fear, joy, sadness, disgust, and surprise function as stable targets for classification.

This categorical alignment is computationally convenient, as fixed emotion labels translate directly into supervised training objectives. However, empirical findings in affective science complicate this assumption. Reviews of physiological and behavioral evidence indicate substantial variability within emotion categories and overlap across categories, challenging the claim that emotions possess distinct biological signatures (Barrett, 2006; Kreibig, 2010; Mauss & Robinson, 2009). Similarly, observational research suggests that prototypical facial configurations are infrequent in natural interaction and highly context-dependent (Barrett, 2011; Russell et al., 2003). Such variability weakens the assumption that discrete categories correspond to stable, universally encoded signals.

In contrast, constructionist accounts conceptualize emotions as dynamically assembled through the integration of bodily states, contextual information, conceptual knowledge, and language (Barrett, 2017). Within this framework, emotion categories function as interpretive constructs rather than biologically

discrete entities. Emotional meaning is therefore context-sensitive and multimodal, emerging from interaction rather than residing solely in linguistic form. Recently, robust empirical evidence has supported constructionist accounts of emotion modelling (Azari et al., 2020).

This theoretical divergence has direct implications for automated emotion recognition. If emotional experience is constructed through multimodal and contextual processes, unimodal systems - particularly those restricted to text - may capture lexical regularities without fully representing affective dynamics (Wu et al., 2025). Classification performance under such constraints may reflect alignment with labelling conventions rather than faithful modelling of emotional states. The tension between categorical convenience and contextual complexity thus forms the conceptual foundation of the present study.

3 Methodology: The Methodological Pipeline

3.1 Dataset and Signal Conversion

The dataset consists of unscripted dialogue extracted from the televised Polish culinary competition *MasterChef Polska*, a high-arousal conversational environment.

The training set consists of 28,395 sentences utilized for training and validation, sourced from a diverse corpus of unscripted television programming - including survival reality shows, competitive tournaments, and spontaneous dialogue - across multiple languages, including but not limited to Dutch, Arabic, Slovak, Polish, and Russian while the target language was English. Despite this multi-program training base, the study maintains a rigorous focus on *MasterChef Polska* as the primary case analysis for evaluation, utilizing a separate test set of 791 sentences (source: Polish; target: English). All instances were analyzed at the sentence level, with each sentence assigned a single dominant emotion label from seven classes: joy, sadness, anger, fear, surprise, disgust, and neutral.

Although the raw material exists as audio-visual recordings, the analytical pipeline intentionally retained only the textual modality to simulate standard text-mining constraints. Visual and acoustic features such as facial expression and prosody were excluded.

3.1.1 Automated Speech-to-Text (STT)

Audio recordings were transcribed using the AssemblyAI Universal 2 model (Khare et al., 2025). Model selection was based on benchmarking against OpenAI Whisper-Large (Radford et al., 2023) using Word Error Rate (WER) (Jurafsky & Martin, 2023). The WER analysis was drawn from a manually verified 15-minute subset of unscripted dialogue to ensure the benchmark reflected the "high-stress" environment of the show. Universal 2 achieved a lower overall WER and provided sentence-level segmentation compatible with downstream emotion classification. Consequently, it was adopted for the final pipeline.

$$WER = \frac{S+I+D}{N} * 100 \quad (1)$$

Equation 1. Word Error Rate (WER), where *S* denotes the total words replaced in transcription by a model, *I* words added by a model that did not match original transcription, *D* total words omitted by a model, *N* number of words in total.

Table 1: Word Error Rate (WER)

Model	Substitutions	Insertions	Deletions	Total Words	WER
Universal 2	88	12	27	1167	10.883%
Whisper-Large	86	25	11	1043	11.697%

3.1.2 Ground Truth Construction

The training set utilized labels inherited from the client's proprietary, third-party LLM-based emotion recognition system. While the specific model architecture remains undisclosed due to commercial sensitivity, the system was confirmed to translate the source Polish text to English to generate the training labels used in this study. The test set (N = 791) was manually annotated by native Polish-speaking students. To ensure consistency, the student team reviewed the manually prepared evaluation subset against the project's specific annotation guidelines provided by the client, cross-referencing utterances with original audio-visual recordings to account for facial expressions and prosodic cues. This methodology results in an imbalanced distribution across the seven emotional classes - joy, sadness, anger, fear, surprise, disgust, and neutral - as detailed in Table 2.

Table 2: Distribution of Emotional Classes Across Datasets

Emotion Class	Training (LLM)	Training (%)	Evaluation (Manual)	Evaluation (%)
Surprise	12,164	42.8	413	52.2
Joy	5,554	19.6	120	15.2
Anger	1,909	6.7	35	4.4
Neutral	3,301	11.6	102	12.9
Sadness	1,867	6.6	85	10.8
Fear	1,800	6.35	27	3.4
Disgust	1,800	6.35	9	1.1
Total	**28,395**	**100**	**791**	**100**

This introduces a labeling asymmetry: the training data reflects automated labeling conventions, whereas the evaluation phase relies on manually verified ground truth. As a result, the classifier is evaluated against a human-verified standard to determine how effectively a model trained on operational AI labels can align with human emotional interpretation. We define this dependency as a "Ground Truth Paradox": while the model is optimized to align with a commercial labeling framework, its ultimate validity is tested against the nuanced benchmarks of native human perception (Bender et al., 2021; Schuller et al., 2026). This limitation is explicitly considered in the interpretation of results.

3.2 Machine Translation

Because the classifier operates in English, Polish transcripts were translated using the MarianMT Polish–English neural machine translation model based on the OPUS-MT framework (Tiedemann & Thottingal, 2020). The pipeline follows a sequential structure:

Audio → Transcription → Translation → Emotion Classification

No multilingual joint training or multimodal fusion was implemented. Translation therefore constitutes an independent transformation stage prior to classification.

3.3 Model Architectures and Benchmarking

To assess the impact of architectural complexity, three model classes were evaluated using identical splits:

- Support Vector Machine (SVM) with linear kernel and term frequency–inverse document frequency (TF-IDF) features (Cortes & Vapnik, 1995; Jurafsky & Martin, 2023),
- Bidirectional LSTM (Hochreiter & Schmidhuber, 1997), the architecture consisted of a 300-dimension embedding layer, two bidirectional LSTM layers with a hidden size of 128, and a dropout rate of 0.3,
- Fine-tuned RoBERTa-base transformer (Liu et al., 2019; Wolf et al., 2020).

All models performed sentence-level single-label classification. Comparative benchmarking allowed assessment of whether performance differences arose from increased contextual modelling capacity. Individual sentences were analyzed as the primary unit of classification to simulate the constraints of real-time media monitoring and to minimize complexity in the multi-speaker environment of a televised cooking competition. While longer contexts might theoretically yield higher classification accuracy by resolving cross-speaker ambiguities, this study focuses on the structural limits of affect recognition from isolated linguistic markers.

3.4 Training Procedure

RoBERTa was fine-tuned using AdamW (Loshchilov & Hutter, 2019) with a class-weighted cross-entropy loss. This approach assigns higher penalties to misclassifications of minority classes, ensuring the model does not become biased toward the majority class. Back-translation augmentation (Sugiyama & Yoshinaga, 2019) was applied to increase exposure to minority classes. Performance was evaluated using accuracy, precision, recall, and F1-score metrics (Jurafsky & Martin, 2023).

3.5 Explainable AI (XAI) Audit

To examine model decision behavior beyond aggregate performance metrics, a post-hoc interpretability method was applied: Gradient × Input (Shrikumar et al., 2017). This method was used to analyze decision patterns in correctly classified and misclassified instances to determine which specific linguistic features drove the model's predictions.

4 Results

The following section presents the quantitative performance metrics of the methodological pipeline, comparing speech-to-text accuracy, baseline model benchmarks, and the specific failure modes identified through Explainable AI (XAI).

4.1 Performance Benchmarking Across Model Generations

The three model classes described in §3.3 were evaluated on the manually annotated human-verified test set (N=791). The SVM baseline, implemented with TF-IDF vectorization and a 50,000-token vocabulary, achieved an overall accuracy of 0.467. The model demonstrated moderate performance on high-frequency and high-arousal categories such as anger and joy, with an average F1-score of 0.547 across these classes. However, it failed to detect any instances of the disgust category during testing. This error indicates a substantial limitation in capturing nuanced affective distinctions using surface-level lexical representations.

The Bidirectional LSTM model, equipped with embedding, pooling, and fully connected layers, yielded an overall accuracy of 0.45. Although precision for the fear category was relatively high, recall remained low at 0.23. This imbalance suggests that while the model successfully identified prototypical expressions of fear, it struggled to generalize more subtle or context-dependent instances present in unscripted dialogue.

The transformer-based RoBERTa model substantially outperformed both baselines, achieving an overall accuracy of 0.755. The performance gain highlights the advantage of contextualized token representations for emotion classification in high-variance conversational text.

It is important to note that the dataset exhibits strong class imbalance, with the surprise category representing more than half of the evaluation instances (413 of 791). This distribution influences overall performance metrics and partially explains the higher weighted F1-score relative to the macro-average.

Table 3: Results Comparison Table (Evaluated on Manually Labeled Test Set, N=791)

Model	Accuracy	Precision	Recall	F1-score
SVM	0.467	0.478	0.343	0.358
Bi-LSTM	0.450	0.570	0.452	0.419
RoBERTa	0.755	0.758	0.755	0.755

4.1.1 The Role of Neural Machine Translation

The move from SVM and Bi-LSTM to RoBERTa improves accuracy to 0.755 on the test set, but the model's performance heavily depends on English-centric pre-training and the quality of translation. Polish transcripts are converted via the MarianMT NMT layer, which can introduce semantic distortions. About 9.7% of classification errors originate in this translation stage, often due to flattened Polish morphological markers or literal translations that misrepresent context. For instance, "Synu" can convey paternal authority or informal camaraderie depending on its suffix, but is usually translated as "Son," losing interpersonal nuance. Similarly, "Granat," meaning pomegranate in MasterChef Polska, is sometimes translated as "Grenade," producing false signals of anger or fear for the classifier. These examples illustrate that even a high-performing model like RoBERTa remains vulnerable to contextual distortions if the translation layer does not preserve cultural and pragmatic cues.

The 9.7 percent translation error was calculated through a manual forensic review of all 194 misclassified instances produced by the RoBERTa model. Two researchers from the student team cross-referenced the model's incorrect predictions with the original Polish audio and the translated English transcripts. An error was attributed to the translation layer if the English text contained a specific semantic distortion - such as the literal rendering of the culinary term "Granat" (pomegranate) as "Grenade" - that logically misled the classifier into a high-intensity emotional trigger like anger or fear.

4.2 Granular Failure Modes

Although the RoBERTa architecture substantially outperformed the SVM and Bi-LSTM baselines, confusion matrix analysis reveals uneven performance across emotional categories. Performance varied significantly depending on class frequency within the dataset.

The disgust category represents the most extreme example of class imbalance, with only nine instances in the evaluation set. Despite this limitation, the model achieved a recall of 0.667 and an F1-score of 0.632, indicating that lexical markers associated with disgust were detected in several cases. However, the small sample size limits the reliability of performance estimates and increases sensitivity to misclassification.

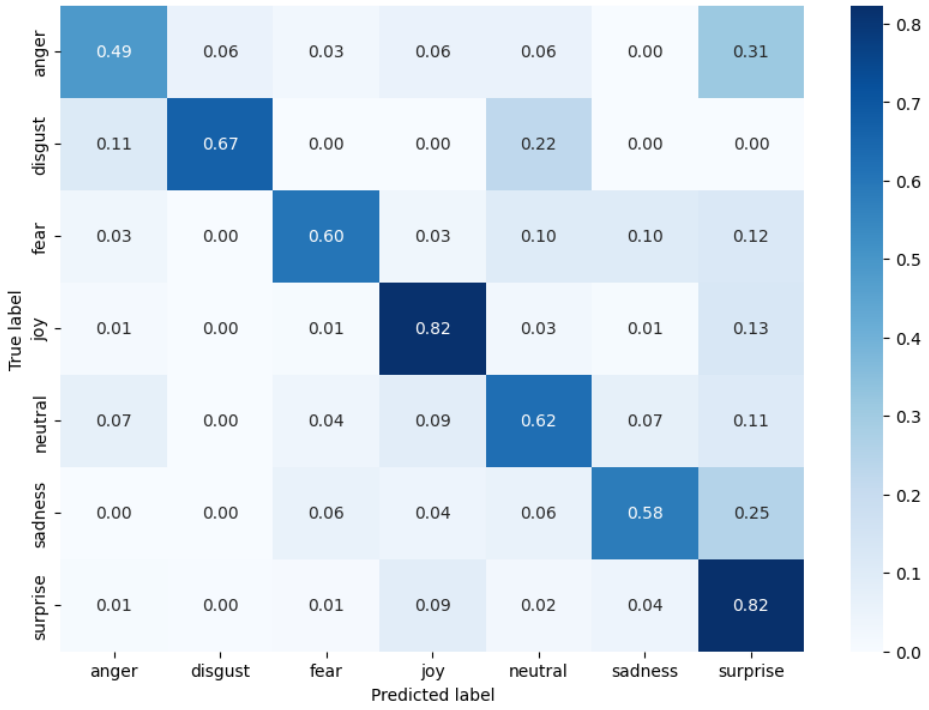

Figure 1: Confusion Matrix for RoBERTa-base (Row = True Label, Column = Predicted Label)

Source: Own

Lower performance was observed for anger, which achieved a recall of 0.486 and an F1-score of 0.531, as well as sadness (F1 = 0.578) and neutral (F1 = 0.588). These results suggest that certain emotional states may be expressed through more subtle linguistic cues, making them harder to capture using textual signals alone.

In contrast, the model achieved strong performance for surprise (F1 = 0.835) and joy (F1 = 0.767), both of which appear frequently in the dataset and are often associated with explicit lexical indicators.

Table 4: Per-Class Metrics Table

Emotion	Precision	Recall	F1-score
Surprise	0.848	0.823	0.835
Joy	0.723	0.817	0.767
Anger	0.586	0.486	0.531
Neutral	0.556	0.625	0.588
Sadness	0.574	0.582	0.578
Fear	0.745	0.603	0.667
Disgust	0.600	0.667	0.632

4.2.1 High-Arousal Confusion Patterns

A recurring confusion pattern was observed between several high-arousal emotional categories. While anger maintained moderate recall (0.486), 18 of the 35 anger instances were misclassified, most frequently as joy or surprise.

Inspection of misclassified examples indicates that the model often responds to intensity-related lexical signals without consistently distinguishing emotional valence. In the absence of acoustic prosody or facial cues, strongly expressive language may be interpreted as high-arousal positivity rather than high-arousal negativity. This suggests that arousal and valence dimensions are not fully disentangled within the textual representation learned by the model.

4.3 Explainable AI Audit

To investigate model behavior beyond aggregate performance metrics, a Gradient × Input analysis was conducted to analyze token-level attribution patterns.

4.3.1 Gradient × Input Attribution

Gradient × Input (Shrikumar et al., 2017) was used to compute the partial derivative of the output with respect to input embeddings, producing token-level relevance scores. Across multiple examples, the model frequently assigned disproportionate weight to isolated, emotionally salient tokens. In several misclassifications, these tokens dominated the prediction despite contradictory contextual information within the same sentence.

To visually represent these attributions, **bold text** is used below to denote tokens with the highest positive attribution weights (those most responsible for the model's chosen label). For example, the sentence "High stakes, you only win by taking **risks**" was labeled anger but predicted as disgust. The mathematical audit indicated that the intensity-related surface-level signal of the word **risks** outweighed the broader competitive context of the statement.

In another case, "I **love** to eat good things" was labeled joy but predicted neutral, illustrating how positive lexical cues like **love** do not always dominate classification when the surrounding contextual signals are structurally weak. A different failure pattern is observed in the sentence "we all **expected** something completely **different**," which was labeled as neutral but predicted as joy. Here, the model appears to overgeneralize and assign emotional meaning to the words **expected** and **different**, suggesting that emotionally suggestive words can trigger classification even in the absence of a clear affective signal. Overall, these token-level attributions suggest that the model relies on sparse lexical cues rather than distributed contextual reasoning, often misclassifying utterances when isolated words conflict with the broader conversational context.

5 Discussion

5.1 The Ground Truth Paradox and Epistemological Validity

The study identifies a critical epistemological risk defined as the "Ground Truth Paradox"**.** Evaluating the classifier against AI-generated labels creates a circular dependency where models are optimized to reproduce the patterns of other models rather than verified emotional states. Even with high accuracy, such systems may

reflect system-level consistency within a specific labeling framework rather than definitive emotional truth. This dependency suggests that future research must prioritize native human-annotated "gold standards" to move beyond automated bias.

5.2 Regulatory Implications: The EU AI Act and "Context Blindness"

The identified 9.7% translation error rate and the model's reliance on isolated lexical cues highlight the fragility of cross-lingual affective pipelines. For organizations deploying these tools under the EU AI Act, particularly in high-risk domains such as recruitment, mental health, or creditworthiness, this "Context Blindness" represents a significant technical and ethical liability.

Because text-only systems are prone to misinterpreting high-arousal states due to a lack of situational "situatedness," their deployment risks automating systemic bias. We conclude that responsible AI development must transition from unimodal text-mining to multimodal, context-aware architectures to bridge the gap between categorical convenience and the complex reality of human affective experience.

5.3 Limitations

Several factors bound the generalizability of this study. First, the evaluation focuses on a single-domain dataset (*MasterChef Polska*), where the emotional repertoire is skewed toward high-arousal interactions. Second, a distinct domain and language shift exists between the training and evaluation phases. While the inclusion of diverse unscripted international programming provided a broad emotional foundation for training, the model was evaluated on a specific Polish-to-English pipeline that was only partially represented during the learning stage. Finally, reliance on a single-sentence unit of analysis prevents the model from utilizing broader conversational flow.

Acknowledgment

The authors would like to express their sincere gratitude to Mekselina Doğanç for her excellent mentorship and valuable guidance throughout the project.

This research article was developed with the assistance of Gemini (Google, Gemini 3 Flash) and ChatGPT (OpenAI, GPT-5). These tools were utilized as a writing aid to assist with adjusting and rephrasing original sentences and providing feedback on the logical structure of the text.

References

Azari, B., Westlin, C., Satpute, A. B., Hutchinson, J. B., Kragel, P. A., Hoemann, K., Khan, Z., Wormwood, J. B., Quigley, K. S., Erdogmus, D., Dy, J., Brooks, D. H., & Barrett, L. F. (2020). Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience. *Scientific Reports*, *10*(1), 20284. https://doi.org/10.1038/s41598-020-77117-8

Barrett, L. F. (2006). Solving the Emotion Paradox: Categorization and the Experience of Emotion. *Personality and Social Psychology Review*, *10*(1), 20–46. https://doi.org/10.1207/s15327957pspr1001_2

Barrett, L. F. (2011). Constructing emotion. *Psihologijske Teme*, *20*(3), 359–380.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23. https://doi.org/10.1093/scan/nsw154

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–623. https://doi.org/10.1145/3442188.3445922

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*(3), 550–553. https://doi.org/10.1037/0033-295x.99.3.550

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Youcanprint.

Khare, Y., Peyash, T., Vanzo, A., & Yoshioka, T. (2025). *Universal-2-TF: Robust All-Neural Text Formatting for ASR* (arXiv:2501.05948). arXiv. https://doi.org/10.48550/arXiv.2501.05948

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology, The Biopsychology of Emotion: Current Theoretical and Empirical Perspectives*, *84*(3), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Loshchilov, I., & Hutter, F. (2019). *Decoupled Weight Decay Regularization* (arXiv:1711.05101). arXiv. https://doi.org/10.48550/arXiv.1711.05101

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, *23*(2), 209–237. https://doi.org/10.1080/02699930802204677

Picard, R. W. (2000). *Affective Computing*. MIT Press.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

Regulation (EU) 2024/1689 (2024). http://data.europa.eu/eli/reg/2024/1689/oj

Russell, J. A., Bachorowski, J.-A., & Fernández-Dols, J.-M. (2003). Facial and Vocal Expressions of Emotion. *Annual Review of Psychology*, *54*(Volume 54, 2003), 329–349. https://doi.org/10.1146/annurev.psych.54.101601.145102

Schuller, B., Mallol-Ragolta, A., Almansa, A. P., Tsangko, I., Amin, M. M., Semertzidou, A., Christ, L., & Amiriparian, S. (2026). Affective computing has changed: The foundation model disruption. *Npj Artificial Intelligence*, *2*(1), 16. https://doi.org/10.1038/s44387-025-00061-3

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *Proceedings of the 34th International Conference on Machine Learning*, 3145–3153. https://proceedings.mlr.press/v70/shrikumar17a.html

Sugiyama, A., & Yoshinaga, N. (2019). Data augmentation using back-translation for context-aware neural machine translation. In A. Popescu-Belis, S. Loáiciga, C. Hardmeier, & D. Xiong (Eds.), *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)* (pp. 35–44). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-6504

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – Building open translation services for the World. In A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, & M. L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 479–480). European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.61/

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. von, Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. https://doi.org/10.48550/arXiv.1910.03771

Wu, Y., Mi, Q., & Gao, T. (2025). A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. *Biomimetics*, *10*(7). https://doi.org/10.3390/biomimetics10070418