# **The Shift in Affective Computing: A Comparative Analysis of Transformer and Statistical Models for Text Emotion Detection Since 2020**

### **Abstract**

Since 2020, the field of text-based emotion detection, a nuanced sub-discipline of sentiment analysis, has undergone a profound paradigm shift. This report provides an exhaustive comparative analysis of the two dominant modeling approaches: traditional statistical methods, such as Support Vector Machines (SVM) and Naive Bayes (NB), and the now-ascendant transformer-based architectures, including BERT, RoBERTa, and their variants. The analysis reveals a decisive performance advantage for transformer models, which consistently achieve state-of-the-art results across a multitude of benchmarks. This superiority is rooted in their fundamental architectural innovation—the self-attention mechanism—which enables an unprecedented understanding of linguistic context, a critical bottleneck that constrained previous feature-dependent statistical models.

This report deconstructs the architectural principles of both paradigms, synthesizes empirical performance data from recent academic literature, and provides a critical survey of the benchmark datasets and evaluation metrics that shape the field. While transformers demonstrate superior accuracy, often exceeding 90% on standard tasks, the analysis highlights a nuanced landscape of trade-offs. The choice of an optimal model is shown to be a strategic decision involving an efficiency-accuracy spectrum, where lighter, distilled models like DistilBERT offer a compelling balance for real-world, resource-constrained applications. Furthermore, the report delves into the advanced frontiers of affective computing, examining how these models contend with complex linguistic phenomena such as sarcasm and irony, the pressing challenge of model interpretability, and the expansion into multilingual and hybrid model configurations. The findings indicate that while the problem of classifying basic emotions in clean text is largely solved, the future of the field lies in addressing these more complex challenges, pushing towards a more holistic, robust, and transparent understanding of human emotion.

## **Section 1: Architectural Foundations of Modern Emotion Detection**

The divergence in performance between statistical and transformer models in text emotion detection is not a matter of simple iteration but a consequence of fundamentally different architectural philosophies. Statistical models operate on an abstraction of language, treating text as a collection of features that must be manually engineered and are largely devoid of context. In stark contrast, transformer models are designed to process language in its native, sequential form, with an architecture built around the core principle of understanding context. Examining these foundational differences is essential to comprehending the performance chasm that has defined the field since 2020.

### **1.1 The Statistical Bedrock: Context-Agnostic Classification**

For decades, statistical models formed the bedrock of text classification, including early efforts in emotion detection. Their approach is characterized by a two-stage process: first, converting unstructured text into a structured, numerical format through feature engineering, and second, applying a mathematical algorithm to classify these features. The two most prominent models in this paradigm are the Support Vector Machine (SVM) and the Naive Bayes (NB) classifier.

#### **Mechanics of Statistical Models**

The enduring presence of Naive Bayes and SVM in comparative studies is a testament to their efficiency and historical importance. They serve as crucial baselines against which the advancements of deep learning are measured.1

The **Naive Bayes (NB)** classifier is a probabilistic model grounded in Bayes' theorem. It calculates the probability that a given text belongs to a specific emotion class based on the probabilities of the individual words appearing in that class.4 Its power lies in its simplicity and computational efficiency, particularly on high-dimensional data like text. However, this efficiency comes from a significant compromise: the "naive" assumption that all features (words) are independent of one another given the class.4 In natural language, where word order and co-occurrence are rich with meaning, this assumption is almost always false. For example, the model considers "not" and "good" as two independent events, failing to capture the negation that inverts the sentiment. Despite this, NB performs surprisingly well, especially when computational resources are limited or datasets are small.4

The **Support Vector Machine (SVM)** classifier, in contrast, is a geometric model. It conceptualizes each text document as a point in a high-dimensional space, where each dimension corresponds to a feature. The algorithm's objective is to find an optimal hyperplane—a decision boundary—that best separates the points of different emotion classes with the maximum possible margin.4 This maximization of the margin is key to its ability to generalize well to unseen data. For complex, non-linear relationships between features, SVMs employ the "kernel trick," using functions like the Radial Basis Function (RBF) to project the data into an even higher-dimensional space where a linear separation becomes possible.6 SVMs have a reputation for high accuracy, especially on smaller, well-curated datasets, and are considered robust against overfitting.4

#### **The Centrality of Feature Engineering**

The performance of any statistical model is inextricably tied to the quality of the features it is given. This process, known as feature engineering, is the critical, and ultimately limiting, step in the statistical NLP pipeline. It is the process of manually selecting, transforming, and creating variables from raw text data to make it suitable for machine learning algorithms.7

The most common techniques for text feature extraction are fundamentally context-agnostic. The **Bag-of-Words (BoW)** model represents a document as an unordered collection—a "bag"—of its words, disregarding grammar and word order entirely. Each document is converted into a numerical vector where each element corresponds to the frequency of a word from a predefined vocabulary.9 A more sophisticated variant is

**Term Frequency-Inverse Document Frequency (TF-IDF)**. This method improves upon BoW by weighting each word's count. It increases the weight for words that are frequent in a specific document but rare across all documents in the corpus, thereby giving more importance to terms that are more discriminative.9 To capture some local word order, these methods can be extended to

**N-grams**, which treat sequences of N consecutive words (e.g., "very good," "not so bad") as single tokens.10

Researchers have also attempted to manually inject semantic information into the feature set. This includes creating features based on sentiment lexicons (lists of words with pre-assigned emotional scores), the presence of emoticons and emojis, and rules to handle negation words like "not" or "never".9

#### **Inherent Architectural Limitations**

The reliance on handcrafted features imposes a hard ceiling on the capabilities of statistical models. Their architectural limitations are not bugs to be fixed but are inherent to their design.

The primary flaw is **context blindness**. Because BoW and TF-IDF discard word order and grammatical structure, the models cannot differentiate between "The movie was not good at all" and "The movie was good, not at all bad." The word "sick" is represented by the same feature value whether it appears in "this performance was sick" (positive) or "I feel sick" (negative).11 This inability to disambiguate words based on their surrounding context is the single greatest weakness of this approach.

Furthermore, the foundational assumptions of these models clash with the reality of language. The feature independence assumption in Naive Bayes is a clear violation of linguistic principles, which can impair performance when complex word relationships are key to determining emotion.4 For SVMs, performance is highly sensitive to the meticulous tuning of hyperparameters like the kernel type and regularization parameters, and their training time becomes computationally prohibitive for the massive datasets common today.4 Ultimately, both models are fundamentally constrained by the information they are given; they cannot learn patterns or relationships that are not explicitly encoded in the feature vectors created by the developer. The transition to a new paradigm was not merely a quest for higher accuracy scores but a necessary response to an architectural dead end. The field had reached the limits of what could be achieved by manually crafting features to represent the immense complexity of human language, compelling a pivot to an architecture that could learn these representations automatically from the data itself.

### **1.2 The Transformer Revolution: Context as a Core Principle**

Introduced in 2017, the Transformer architecture and the models derived from it, such as BERT, marked a revolutionary departure from prior approaches. Instead of relying on human-engineered features, these models learn rich, contextual representations of language directly from vast amounts of text data. This is achieved through two key innovations: the self-attention mechanism and the pre-train/fine-tune paradigm.

#### **The Self-Attention Mechanism**

At the heart of the Transformer is the **self-attention mechanism**. This mechanism addresses the critical limitation of context blindness that plagued earlier models. When processing a sentence, self-attention allows the model to dynamically weigh the importance of all other words in the input sequence for understanding each individual word.11 For example, in the sentence "The customer was angry because the package arrived broken," when encoding the word "angry," the self-attention mechanism can learn to pay high attention to "broken," establishing a clear causal link that defines the emotional context.

This process is accomplished by creating three vector representations for each word: a Query, a Key, and a Value. The Query vector of a given word is matched against the Key vectors of all other words in the sequence to compute attention scores. These scores, normalized via a softmax function, determine how much of each word's Value vector should be factored into the final representation of the original word.13 This allows the model to solve the long-range dependency problem—understanding relationships between words far apart in a text—which was a significant challenge for previous sequential models like Recurrent Neural Networks (RNNs).14 The ability of models like BERT to consider both the left and right context simultaneously (i.e., bidirectionality) is a direct result of this mechanism and is the source of their deep contextual understanding.16

#### **The Pre-train and Fine-tune Paradigm**

The second transformative innovation is the two-phase learning strategy.

1. **Pre-training:** In the first phase, a transformer model is pre-trained on an enormous corpus of unlabeled text, such as the entirety of English Wikipedia and the BooksCorpus (totaling 16GB for the original BERT).12 The training is "self-supervised," meaning the model learns from the data itself without human-provided labels. This is done through objectives like  
   **Masked Language Modeling (MLM)**, where the model learns to predict randomly masked words in a sentence based on the surrounding unmasked words. For example, in "The customer felt about the service," the model learns to predict words like "happy," "angry," or "indifferent" based on the context. Another original objective was **Next Sentence Prediction (NSP)**, a binary classification task to determine if two sentences appeared consecutively in the original text, intended to help the model understand sentence relationships.15 This pre-training phase imbues the model with a profound, generalized understanding of grammar, syntax, semantics, and world knowledge.
2. **Fine-tuning:** In the second phase, the pre-trained model is adapted for a specific downstream task, such as emotion detection. This is done by adding a small, task-specific classification layer on top of the pre-trained model and training it on a much smaller, labeled dataset (e.g., a few thousand sentences annotated with emotions). During fine-tuning, the powerful, pre-trained representations are leveraged, and the model's weights are only slightly adjusted to specialize in the target task. This transfer learning approach is incredibly effective and data-efficient, allowing for state-of-the-art performance without needing to train a massive model from scratch for every new task.15

#### **Anatomy of Key Models (Since 2020)**

While BERT laid the groundwork, the post-2020 landscape is dominated by its refined successors.

**BERT (Bidirectional Encoder Representations from Transformers):** The foundational model, introduced by Google in 2018, remains a central point of comparison. Its key contribution was demonstrating the power of deep bidirectionality through the MLM and NSP pre-training tasks.15

**RoBERTa (Robustly Optimized BERT Approach):** Introduced by Facebook AI in 2019, RoBERTa is not a new architecture but a set of optimized training procedures for BERT that yield significantly more powerful models.19 The evolution from BERT to RoBERTa reflects a maturing understanding within the research community of what constitutes effective pre-training. The focus shifted from adding new architectural components to optimizing the training recipe itself. The key changes were:

* **Removal of the NSP objective:** The RoBERTa authors found that removing the Next Sentence Prediction task, a core component of BERT's pre-training, slightly improved performance on downstream tasks, suggesting it was less beneficial than originally thought.16 This finding demonstrated a crucial refinement in understanding what pre-training objectives are truly valuable.
* **Massively increased training data:** RoBERTa was trained on 160GB of text, over ten times more than BERT's 16GB.19
* **Larger batch sizes and longer sequences:** The model was trained with much larger mini-batches, which was found to improve both the training objective and final task accuracy.19
* **Dynamic masking:** In BERT's pre-training, the masking pattern for sentences was generated once during data preprocessing and remained static. RoBERTa introduced dynamic masking, where a new masking pattern is generated every time a sequence is fed to the model, improving the variety of the training data and the model's robustness.16  
    
  These optimizations collectively result in RoBERTa consistently outperforming BERT on a wide range of NLP benchmarks, including emotion detection.16

**DistilBERT:** Recognizing the immense computational cost of models like BERT, researchers at Hugging Face developed DistilBERT, a smaller, faster, and lighter version.12 It is created using a technique called

**knowledge distillation**. In this process, a smaller "student" model (DistilBERT) is trained to mimic the output probabilities of a larger, pre-trained "teacher" model (BERT). The student learns to replicate the rich behavior of the teacher, effectively compressing its knowledge. The result is a model that is 40% smaller and 60% faster than BERT while retaining approximately 97% of its language understanding capabilities, making it an excellent choice for applications on mobile devices or in real-time, latency-sensitive environments.15

## **Section 2: A Synthesis of Empirical Performance**

While architectural principles explain the theoretical potential of different models, empirical results from comparative studies provide the definitive verdict on their practical efficacy. Since 2020, a wealth of research has benchmarked statistical, deep learning, and transformer-based models against one another on a variety of emotion detection tasks. The collective evidence paints a clear picture: transformers represent a significant leap forward in performance, though this advantage must be weighed against their computational demands.

### **2.1 Quantitative Benchmarking on Standardized Tasks**

Across numerous studies, datasets, and languages, a consistent performance hierarchy has emerged. Transformer-based models decisively outperform traditional machine learning methods and earlier deep learning architectures like LSTMs and RNNs.14 While statistical models typically achieve accuracies in the 70-85% range, transformers frequently surpass the 90% mark on similar tasks.14

Statistical models, while outmatched, serve as essential baselines. In direct comparisons, SVMs generally exhibit higher accuracy than Naive Bayes.4 For instance, a 2022 study on emotion detection in tweets found that an SVM model achieved 84% accuracy, which was superior to Naive Bayes classifiers but was in turn surpassed by transformer models. The same study showed a BERTweet model (a BERT variant trained on Twitter data) reaching 89% accuracy, and a novel ensemble model combining BERT and SVM achieving an even higher 91%.6 This trend holds across different languages and data sources. A 2024 study on Indonesian-language Twitter data reported BERT achieving 75.6% accuracy, comfortably ahead of SVM at 71.6% and MultinomialNB at 64.8%.24 Similarly, research on Turkish text found that systems based on BERT and ELECTRA (another transformer variant) achieved accuracies of 70% and 72% respectively when paired with classical machine learning classifiers for final prediction.25

Within the transformer family itself, a nuanced hierarchy is evident. As a result of its optimized pre-training, **RoBERTa** consistently matches or exceeds the performance of the original BERT model.15 One comparative analysis on the formal ISEAR dataset identified RoBERTa as the top-performing model with an accuracy of 74.3%, followed by XLNet (72.9%), BERT (70.0%), and DistilBERT (66.9%).26 Another study from 2024 highlighted a

twitter-roberta-base model, specifically pre-trained on Twitter data, achieving an impressive 92% accuracy on an emotion classification task, outperforming other transformer variants.12

However, the narrative that "bigger is always better" is compellingly challenged by certain findings. In a comprehensive 2024 study that used the large GoEmotions dataset (with emotions consolidated into three classes: positive, negative, neutral), the smaller, more efficient **DistilBERT** model achieved a remarkable accuracy of 95.88%. This result was not only the highest but also significantly surpassed its much larger and more complex counterparts, including RoBERTa (86.56%), XLNet (82.34%), BERT (78.52%), and even the massive GPT-3.5 model (68.35%).27 This unexpected outcome suggests that for certain task formulations and data distributions, a well-distilled, compact model can be more effective than a larger, more general-purpose one. This finding moves the field beyond a monolithic "leaderboard-chasing" mentality. It demonstrates that the optimal model choice is not absolute but is a function of multiple variables, including the specific emotion taxonomy, the dataset's characteristics, and the application's operational constraints. This validates the need for a diverse ecosystem of models and elevates model selection from a simple performance comparison to a strategic, engineering-driven decision.

The following table synthesizes performance metrics from various studies since 2020, providing a direct quantitative comparison across model paradigms and datasets.

**Table 1: Comparative Performance Metrics of Statistical vs. Transformer Models on Key Emotion Datasets**

| Model | Dataset | Key Metric (Accuracy %) | Secondary Metric (Macro F1-Score) | Source(s) |
| --- | --- | --- | --- | --- |
| **Statistical Models** |  |  |  |  |
| SVM | Twitter-based | 92.0 | - | 4 |
| Naive Bayes | Twitter-based | 83.0 | - | 4 |
| SVM | Twitter (5 emotions) | 84.0 | - | 6 |
| SVM | Garuda Indonesia Tweets | 71.6 | 0.50 | 24 |
| MultinomialNB | Garuda Indonesia Tweets | 64.8 | 0.54 | 24 |
| SVM | GoEmotions (mapped to 6) | 63.0 | 0.54 | 28 |
| **Transformer Models** |  |  |  |  |
| BERT-base | Garuda Indonesia Tweets | 75.6 | 0.58 | 24 |
| BERT-base | ISEAR | 70.0 | 0.702 | 26 |
| BERTweet | Twitter (5 emotions) | 89.0 | - | 6 |
| RoBERTa-base | ISEAR | 74.3 | 0.742 | 26 |
| RoBERTa-base | GoEmotions (3 classes) | 86.56 | - | 27 |
| twitter-roberta-base | Emotion (4 classes) | 92.0 | - | 12 |
| DistilBERT | GoEmotions (3 classes) | **95.88** | - | 27 |
| DistilBERT | ISEAR | 66.9 | 0.693 | 26 |
| **Hybrid/Ensemble Models** |  |  |  |  |
| BERT-SVM Ensemble | Twitter (5 emotions) | 91.0 | - | 6 |
| RoBERTa-SVM Hybrid | Sentiment Classification | 92.0 | 0.82 (F1) | 30 |
| Stacking Classifier (ML) | GoEmotions (mapped to 6) | 64.0+ | - | 28 |

*Note: Direct comparison between all results should be done with caution, as datasets, emotion taxonomies, and evaluation protocols vary between studies. The table aims to illustrate general performance trends.*

### **2.2 The Efficiency-Accuracy Spectrum**

The superior accuracy of transformer models is not without its costs. A critical dimension of comparison is the trade-off between performance and computational efficiency, a factor that heavily influences a model's suitability for real-world deployment.

While training statistical models like SVM can be computationally intensive on very large datasets 4, the resources required to pre-train and fine-tune large transformer models are on another order of magnitude.21 This high cost is a significant barrier to entry for researchers and organizations with limited computational budgets.

Beyond training, **inference latency**—the time it takes for a trained model to make a prediction on a new piece of data—is a crucial metric for practical applications. For use cases like interactive chatbots or real-time analysis of social media feeds, low latency is non-negotiable. Here, the trade-off between accuracy and speed becomes stark. A 2025 study provides a clear example: while a RoBERTa model delivered the highest accuracy, its average inference time was 14.7 milliseconds per sample. In contrast, the smaller DistilBERT model was nearly twice as fast, with an average inference time of 8.3 milliseconds.33 This makes DistilBERT a far more viable choice for applications where immediate predictions are required.

The strong performance and high efficiency of models like DistilBERT underscore the immense practical value of knowledge distillation.15 This technique provides a clear pathway for democratizing access to powerful NLP models, enabling their deployment on less powerful hardware or in environments with strict latency constraints. Research has even established a rough hierarchy of computational expense among common transformers, with the order of decreasing cost and time being approximately XLNet, BERT, RoBERTa, and finally, the most efficient, DistilBERT.26 This creates a clear spectrum where practitioners can select a model that aligns with their specific balance of accuracy requirements and resource limitations.

The data suggests that for well-structured, formal text, the performance gains from using larger and more complex transformer models begin to diminish. One study noted that on the formal ISEAR dataset, the accuracy difference between various transformer models was less pronounced, with all of them performing above 88%.33 However, on noisy, informal, short-text data from social media, or in complex conversational settings, the performance differences become more significant.17 This indicates that while the problem of basic emotion classification in "clean" text is nearing a solved state for current architectures, the true frontiers for improvement lie in handling the messiness and complexity of real-world language. The most valuable future research will likely focus not on eking out marginal gains on existing clean benchmarks, but on creating more challenging datasets and developing novel techniques to tackle ambiguity, conversational dynamics, and the subtle, confusable emotion categories that still challenge even the best models.36

## **Section 3: The Ecosystem of Emotion Analysis: Datasets and Metrics**

The performance of any emotion detection model is not an absolute measure but is deeply intertwined with the ecosystem in which it is developed and evaluated. The choice of training dataset fundamentally shapes a model's capabilities, while the selection of evaluation metrics dictates how its performance is perceived and measured. A critical analysis of this ecosystem is necessary to properly contextualize the performance benchmarks presented in the previous section and to understand the subtle biases and priorities that guide research in the field.

### **3.1 The Influence of Benchmark Datasets**

The characteristics of the datasets used for training and testing—their source, size, linguistic style, and annotation schema—have a profound impact on model behavior and reported outcomes.

#### **Dataset Diversity and Characteristics**

The datasets referenced in post-2020 research are diverse, each presenting unique challenges.

* **Social Media Datasets:** A significant portion of modern research leverages data from social media platforms. The **GoEmotions** dataset, with its 58,000 Reddit comments annotated for 27 fine-grained emotions plus neutral, is a prominent example.27 The  
  **MELD (Multimodal EmotionLines Dataset)**, sourced from dialogues in the *Friends* TV series, provides text along with audio and visual data, making it a key resource for conversational and multimodal analysis.34 Other benchmarks like  
  **TweetEval** 39 and various custom-collated Twitter datasets 3 are also common. These datasets are characterized by their informal, idiosyncratic language, including slang, misspellings, abbreviations, and heavy use of emojis. The text is often short and highly context-dependent, posing a significant challenge for models trained on more formal language.17
* **Formal Datasets:** In contrast, the **ISEAR (International Survey on Emotion Antecedents and Reactions)** dataset contains more formally structured, grammatically correct sentences describing emotional events. On this type of data, transformer models tend to achieve higher and less varied performance scores, suggesting the task is less challenging.26
* **Conversational Datasets:** Beyond MELD, datasets like **DailyDialog** 38 are crucial for the sub-field of  
  **Emotion Recognition in Conversation (ERC)**. ERC requires models to go beyond single utterances and model speaker-sensitive dependencies and the flow of emotion over a multi-turn dialogue.36
* **Multimodal Datasets:** While most research still focuses on text-only emotion detection 2, the existence of multimodal datasets like MELD and the newly introduced  
  **SAMSEMO** (with data in five languages) 41 signals a clear future direction. These datasets, which include audio and visual modalities, enable a more holistic analysis of emotional expression, though studies have shown that simply adding more modalities does not always improve performance and can sometimes introduce noise.42

#### **The Impact of Labeling Schema and Imbalance**

The definition of the task itself is determined by the dataset's annotation framework.

* **Emotion Models:** The choice of emotion taxonomy is fundamental. Many studies adopt a **discrete model** based on Paul Ekman's six "basic" emotions: joy, sadness, anger, fear, surprise, and disgust.37 Others use more complex frameworks like Plutchik's wheel of emotions or the highly granular 28-category schema of the GoEmotions dataset.27 The difficulty of the classification task increases dramatically with the number and subtlety of the emotion categories.
* **Class Imbalance:** This is a persistent and critical challenge in nearly all real-world emotion datasets. Emotions are not expressed with equal frequency. Categories like 'joy,' 'sadness,' and 'neutral' are often far more prevalent than 'fear,' 'disgust,' or 'surprise'.12 This imbalance poses a significant risk: a model can achieve high overall accuracy simply by learning to predict the majority classes while completely failing to identify the rarer, but often more critical, emotions. This makes accuracy a potentially deceptive metric.43

The following table provides a profile of some of the major datasets used in recent text emotion detection research, highlighting their key characteristics.

**Table 2: Profile of Major Text Emotion Detection Datasets**

| Dataset Name | Primary Source | Modality | Size (approx.) | Label Taxonomy | Key Characteristics & Use Cases | Source(s) |
| --- | --- | --- | --- | --- | --- | --- |
| **GoEmotions** | Reddit | Text | 58,000 comments | 27 emotions + Neutral | Fine-grained, informal text, class imbalance. Used for multi-class classification. | 27 |
| **Emotion** | General Text | Text | 20,000 texts | 6 emotions (sadness, joy, love, anger, fear, surprise) | Standard benchmark for multi-class text classification. | 44 |
| **ISEAR** | Self-reports | Text | 7,500+ sentences | 7 emotions (joy, fear, anger, sadness, disgust, shame, guilt) | Formal, structured language. Good for testing models on clean text. | 26 |
| **MELD** | *Friends* TV Series | Multimodal (Text, Audio, Video) | 13,000+ utterances | 7 emotions + Sentiment | Conversational, multi-speaker dialogue. Key for Emotion Recognition in Conversation (ERC). | 34 |
| **TweetEval** | Twitter | Text | Multiple subsets | 7 tasks, including Emotion | Standardized evaluation framework for Twitter-specific NLP tasks. | 39 |
| **DailyDialog** | Written Dialogues | Text | 13,000+ dialogues | Annotated for emotion and dialogue acts | High-quality, open-domain conversational data. | 38 |
| **SAMSEMO** | Various Videos | Multimodal (Text, Audio, Video) | 23,000+ video scenes | Ekman's 6 + Neutral/Other | New, multilingual dataset (EN, DE, ES, PL, KO) for MER. | 41 |

A critical examination of these benchmarks reveals a potential pitfall in the field's evaluation methodology. A study on sarcasm detection provided a crucial warning: a model achieved near-perfect accuracy not by understanding sarcasm, but by learning to distinguish the unique writing styles of its two data sources (the satirical *The Onion* versus *The Huffington Post*).46 This was not true sarcasm detection, but "domain detection." This risk is directly applicable to emotion detection. Models that perform exceptionally well on platform-specific datasets like GoEmotions (Reddit) or TweetEval (Twitter) may be learning to recognize the linguistic quirks, slang, and topic distributions of those platforms as much as they are learning abstract emotional expression. A

twitter-roberta-base model's high performance 12 could partially stem from its ability to recognize "this sounds like a tweet labeled 'joy'" rather than from a pure, generalizable understanding of joy. This calls the out-of-domain generalization of many state-of-the-art models into question and highlights an urgent need for more diverse, cross-domain benchmark datasets to ensure researchers are building truly robust emotion classifiers, not just sophisticated platform detectors.47

### **3.2 The Nuances of Model Evaluation**

Just as the choice of dataset defines the problem, the choice of evaluation metric defines what constitutes a "good" solution. Relying on a single, simplistic metric can be misleading and can obscure critical model failures.

#### **Beyond Accuracy**

**Accuracy**, the percentage of correct predictions, is the most intuitive metric but is also the most dangerous, especially in the context of imbalanced datasets.43 If 90% of comments in a dataset are 'neutral,' a model that predicts 'neutral' for every input will achieve 90% accuracy while being completely useless for detecting any actual emotion. For this reason, more nuanced metrics are essential.

#### **The Importance of F1-Score, Precision, and Recall**

A more complete picture of performance is provided by a trio of metrics derived from the model's predictions on a per-class basis:

* **Precision** measures the proportion of true positive predictions among all instances the model predicted as positive (TP/(TP+FP)). High precision is critical when the cost of a false positive is high. In a customer service context, it means not incorrectly flagging a happy customer as angry.1
* **Recall** (also known as sensitivity or true positive rate) measures the proportion of actual positive instances that the model correctly identified (TP/(TP+FN)). High recall is vital when the cost of a false negative is high. In a mental health monitoring application, it means successfully identifying as many individuals expressing distress as possible.1
* **F1-Score** is the harmonic mean of precision and recall (2×(Precision×Recall)/(Precision+Recall)). It provides a single, balanced measure of a model's performance, making it the most common and appropriate metric for tasks with class imbalance.1

#### **Macro vs. Micro vs. Weighted Averaging**

In a multi-class problem like emotion detection, these per-class metrics must be aggregated into a single score. The method of aggregation is a critical choice.

* **Micro-Averaging:** This method calculates the overall metric by summing the individual true positives, false negatives, and false positives across all classes before computing the final score. The result is that the metric is dominated by the model's performance on the most populous classes.49
* **Macro-Averaging:** This method calculates the metric independently for each class and then computes the unweighted average of these scores. This treats every class as equally important, regardless of its frequency in the dataset. It is the most suitable metric for imbalanced classification because poor performance on a rare class will significantly penalize the overall macro-average score.27
* **Weighted-Averaging:** This method also calculates the metric for each class but takes a weighted average, where the weight for each class's score is its proportion in the dataset.

The **Confusion Matrix** is another indispensable diagnostic tool. It is a table that visualizes the performance of a classifier, showing the number of correct and incorrect predictions for each class. It provides a detailed breakdown of errors, revealing which emotions are systematically confused with one another, such as anger and disgust, or fear and surprise.27

The choice between these metrics, particularly macro-F1 versus micro-F1, is not merely a technical detail; it can be interpreted as an ethical stance. In a high-stakes application like monitoring social media for signs of mental distress 14, emotions like 'fear' or 'sadness' are likely to be minority classes. If a research team or developer chooses to optimize for and report a micro-F1 score, they are implicitly prioritizing the model's performance on common emotions like 'joy' or 'neutral.' A high micro-F1 score could be achieved even if the model is completely inept at detecting the rare but critical signals of distress. Conversely, choosing to optimize for a macro-F1 score represents a conscious decision to prioritize the model's ability to identify every emotional state equally, including those of vulnerable individuals. For the responsible development and deployment of AI in sensitive domains, the justification of the chosen evaluation metric should be a mandatory consideration, defended not just on technical grounds but on ethical and application-specific ones.

## **Section 4: Frontiers and Advanced Challenges in Affective Computing**

While transformer models have achieved remarkable success in classifying basic emotions in text, they are still confronted by the deeper complexities of human language and the inherent opacity of their own decision-making processes. The current frontiers of research are focused on pushing beyond simple classification to tackle these advanced challenges, including the interpretation of figurative language, the quest for model transparency, and the expansion of emotion detection into more diverse linguistic and modeling contexts.

### **4.1 The Intricacies of Human Language: Sarcasm and Irony**

One of the most significant hurdles for any text analysis system is the prevalence of non-literal language. Sarcasm and irony, in particular, pose a fundamental challenge because their intended meaning is often the opposite of their literal meaning.51 A comment like "Oh, great, another meeting" does not express the emotion of joy. Accurately detecting this incongruity is a critical prerequisite for correct emotion analysis; failure to do so results in a complete misinterpretation of the user's affective state.53

While transformers, with their deep contextual understanding, are better equipped to handle this nuance than previous models, they still struggle when relying on text alone.53 Research since 2020 indicates that successful sarcasm detection requires moving beyond the utterance itself to incorporate a wider array of contextual signals.

* **Leveraging Context and Metadata:** The most effective approaches integrate rich contextual information into the model's input. This can include the preceding utterances in a conversation, information about the speaker, or external metadata associated with the text.51 For example, a model's ability to detect sarcasm in a news headline is dramatically improved if it knows the headline is sourced from a satirical publication like  
  *The Onion*.46 In conversational data, preserving the thread structure and identifying sentiment shifts between speakers are crucial cues.51
* **Feature Importance:** Studies analyzing the features that contribute most to successful sarcasm detection have found that this external metadata and contextual summarization are far more influential than traditional linguistic features like Part-of-Speech (POS) tags or TF-IDF scores.51 This reinforces the idea that sarcasm is a phenomenon that cannot be understood in isolation.
* **Specialized Models:** The complexity of this problem has led to the development of specialized sarcasm detection models, often using RoBERTa or DistilBERT as a base, but with a sophisticated preprocessing pipeline designed to engineer and integrate these vital contextual features.51

### **4.2 Toward Transparent AI: The Quest for Interpretability**

A major barrier to the widespread adoption of transformers in high-stakes applications—such as mental health assessment or content moderation—is their nature as "black boxes".54 With architectures containing hundreds of millions or even billions of parameters, it is exceedingly difficult to understand

*why* a model made a specific prediction. This lack of transparency erodes trust and makes it challenging to diagnose and correct model failures.54 Consequently, a vibrant area of research has emerged focused on developing techniques for model interpretability.

* **Attention-Based Methods:** A natural and computationally inexpensive approach is to visualize the model's own self-attention scores. The assumption is that these scores, which represent how much "attention" a word pays to other words, can serve as a saliency map, highlighting the parts of the input text that were most influential in the model's decision.13 However, the faithfulness of raw attention scores as a direct explanation of model reasoning has been a subject of debate in the research community.54 More advanced techniques like OPTIMUS have been proposed to systematically evaluate different methods of aggregating attention scores across layers and heads to find the most faithful interpretation for a given prediction, with results showing this approach can be competitive with more complex methods.54
* **Gradient- and Perturbation-Based Methods:** Techniques from the broader machine learning interpretability field, such as LIME (Local Interpretable Model-agnostic Explanations) and Integrated Gradients (IG), can be adapted to transformers. These methods work by systematically perturbing the input (e.g., removing words) or analyzing the gradients of the output with respect to the input features to determine their importance.54
* **Activation-Based Methods:** More recent approaches, like the proposed Contrast-CAT method, aim to improve interpretability by analyzing the model's internal activations. Contrast-CAT works by contrasting the activations generated by an input with a set of reference activations to filter out class-irrelevant features, theoretically producing a cleaner and more faithful attribution map.56

The push for greater performance in handling complex nuances like sarcasm is in direct tension with the goal of simple interpretability. To accurately detect sarcasm, models must ingest and synthesize an ever-wider array of contextual information, including conversational history and user metadata.51 This makes the model's reasoning process inherently more distributed and complex. In contrast, many interpretability methods aim to simplify the explanation by attributing the decision to a few key input tokens.13 A fundamental conflict arises: if a decision truly depends on a subtle interplay of many factors, an explanation that highlights only a few words is necessarily incomplete and potentially misleading. This suggests that as models become more powerful by leveraging holistic context, they may become less amenable to simple, local explanations. The future of interpretability may need to evolve beyond feature-attribution heatmaps toward methods that can generate more holistic, human-readable explanations, such as natural language justifications or reasoning based on prototypical examples.52

### **4.3 Expanding the Horizon: Multilingual and Hybrid Approaches**

To broaden the applicability and push the performance boundaries of emotion detection, researchers are exploring two key frontiers: extending capabilities to more languages and combining different model architectures.

* **Multilingual Emotion Detection:** The vast majority of datasets and research in emotion detection remains English-centric, yet the need for this technology is global.41 Addressing this gap requires models capable of understanding many languages.
  + **Models:** Cross-lingual transformer models like **mBERT** (multilingual BERT) and **XLM-RoBERTa (XLM-R)** are designed for this purpose. They are pre-trained on a corpus containing text from over 100 languages, allowing them to be fine-tuned for emotion detection in languages where labeled data is scarce.41 Studies have shown XLM-R to be particularly effective, outperforming mBERT and other models in multilingual sentiment tasks.57
  + **Challenges:** Significant hurdles remain. Models must contend with **code-switching** (the practice of mixing multiple languages within a single text), a common phenomenon on social media. Furthermore, data scarcity for many languages remains a major obstacle, and performance can still be unsatisfactory. For example, recent studies on Arabic emotion classification show even state-of-the-art models struggling to achieve high accuracy.58 Finally, emotional expression itself can have cultural nuances that a single global model may struggle to capture.
* **Hybrid and Ensemble Models:** In the pursuit of higher accuracy, researchers are finding success by combining the strengths of different architectures.
  + **RoBERTa-SVM:** One innovative approach uses the powerful RoBERTa model not as an end-to-end classifier, but as a sophisticated feature extractor. The rich, contextualized embedding for a text is generated by RoBERTa and then fed into a classical SVM for the final classification step. This hybrid approach has been shown to outperform both the standalone RoBERTa model (with its simple linear classification head) and a standalone SVM (with traditional features).30
  + **Ensemble Models:** Other studies have demonstrated the power of ensembling. One created an ensemble of a BERTweet model and an SVM, which achieved a state-of-the-art 91% accuracy, showing that the two models effectively compensate for each other's weaknesses.6 Another successful approach involved  
    **stacking**, where the predictions of several base machine learning models (e.g., Random Forest, XGBoost, SVM) are used as input for a final "meta-classifier" (e.g., Logistic Regression). On the GoEmotions dataset, such a stacking classifier was found to outperform all individual machine learning models, other ensembles, and even a pre-trained EmoBERTa model.28

The success of these hybrid models, particularly the RoBERTa-SVM configuration, suggests an architectural decoupling of the two primary tasks in classification: "representation learning" and "classification." It implies that while transformers are unequivocally state-of-the-art for the first task—creating a rich, numerical representation of text—classical machine learning algorithms may still offer more powerful or robust methods for the second task—drawing a decision boundary to separate classes in that representation space. This opens a promising new avenue for research in modular, hybrid model design, where practitioners can pair a transformer-based "representation engine" with the optimal "classification engine" for their specific data characteristics.

## **Section 5: Strategic Synthesis and Future Trajectories**

The body of research published since 2020 paints a clear and compelling narrative of progress and transformation in the field of text-based emotion detection. The shift from feature-dependent statistical models to context-aware transformer architectures is complete and decisive. However, the landscape is not monolithic; it is a complex space of trade-offs, emerging challenges, and exciting future directions. Distilling the preceding analysis into a strategic framework can provide actionable guidance for practitioners and illuminate the most promising paths for future research.

### **5.1 A Practitioner's Framework for Model Selection**

The choice of an emotion detection model is not a one-size-fits-all decision. The optimal selection depends on a careful balancing of application requirements, resource constraints, and data characteristics. Based on the empirical evidence, the following framework can guide this decision-making process:

* **For Maximum Accuracy in Offline Analysis:** When the highest possible accuracy is the primary goal and computational resources or latency are not major constraints (e.g., for academic research, market research reports, or deep analysis of customer feedback archives), a **RoBERTa-based model** is the leading choice. Performance can be further enhanced by employing a **hybrid architecture**, such as using RoBERTa as a feature extractor for an SVM classifier, or by using it as a component in a sophisticated **stacking ensemble**.18
* **For Real-Time, Resource-Constrained Applications:** When efficiency is critical, such as in production systems for interactive chatbots, real-time social media monitoring, or on-device applications, **DistilBERT** offers the best-in-class balance of accuracy and speed. It provides performance that is close to its larger counterparts but with significantly lower inference latency and a smaller memory footprint, making it the most practical choice for deployment at scale.15
* **For Baselines or Small-Scale Projects:** When developing a new system or working with smaller datasets, traditional machine learning models like **SVM** can still be highly effective. They are computationally cheaper to train than transformers and provide a robust performance baseline against which the added value of more complex models can be quantitatively measured.1
* **For Multilingual Contexts:** When the task involves non-English text or a mix of languages, a dedicated cross-lingual model is necessary. **XLM-RoBERTa (XLM-R)** is the recommended starting point, as it has demonstrated superior performance over other multilingual models like mBERT. However, practitioners should be aware that significant challenges related to data scarcity and cultural nuance remain, and performance may not match that of English-specific models.57
* **For Tasks with High Linguistic Nuance:** If the application must handle complex phenomena like sarcasm and irony, a transformer model is a prerequisite. However, a standard fine-tuned model is unlikely to be sufficient. The solution requires a more comprehensive pipeline that explicitly engineers and integrates **external context and metadata** to help the model disambiguate non-literal language.51

### **5.2 Uncharted Territories and Future Research Directions**

While the progress has been immense, several key challenges remain, defining the research frontiers for the coming years.

* **Multimodal Emotion Recognition:** The most significant frontier is the move beyond text alone. Human emotion is inherently multimodal, expressed through a combination of words, vocal tone, facial expressions, and gestures. The future of affective computing lies in the ability to fuse these signals. This will involve developing architectures that can effectively combine features from text-based transformers with representations from audio models (like wav2vec 2.0) and computer vision models, leveraging rich, multimodally-aligned datasets like MELD and SAMSEMO.41
* **The Role of Large Language Models (LLMs):** The recent emergence of massive generative models like GPT-3.5, Gemini, and LLaMA introduces a new paradigm for emotion detection. Their potential for **few-shot and zero-shot learning** could enable emotion classification with little to no task-specific fine-tuning, dramatically lowering the barrier for creating new applications.27 However, initial research has shown that smaller, expertly fine-tuned models can still outperform these giants on specific classification tasks.27 Future work will need to explore how to best leverage the vast world knowledge of LLMs while retaining the specialized accuracy of fine-tuned models.
* **Advanced Contextual and Conversational Understanding:** The sub-field of Emotion Recognition in Conversation (ERC) remains a key challenge. Progress will require models that can better track long-range dependencies across conversational turns, model the influence of different speakers, and understand how topics shift and evolve. Methodologies that move beyond simple one-hot emotion labels to capture the reality of **mixed or blended emotions**, such as the proposed Emotion Label Refinement (EmoLR) technique, represent a promising direction toward more nuanced conversational AI.36
* **Better Datasets and Evaluation Protocols:** The field is in constant need of larger, more diverse, and more challenging datasets. Future benchmarks must be explicitly designed to test for out-of-domain generalization. There is a pressing need for more high-quality, annotated datasets in non-English languages, as well as datasets that are multimodally aligned and capture the full spectrum of human emotional expression, including subtle, ambiguous, and culturally specific manifestations.20

### **Conclusion**

The period since 2020 has been transformative for text-based emotion detection. The evidence overwhelmingly confirms a paradigm shift away from the feature-dependent, context-agnostic statistical models of the past and toward the deep, context-aware architectures of transformers. Models like BERT and its more robust successor, RoBERTa, have established a new state-of-the-art, driven by their unparalleled ability to learn the intricate patterns of human language from vast datasets.

However, this analysis reveals that the landscape is far from simple. The dominance of large models is nuanced by a critical trade-off between accuracy and efficiency, where smaller, distilled models like DistilBERT emerge as powerful and practical alternatives for real-world deployment. Furthermore, model performance is shown to be deeply contingent on the nature of the training data and the rigor of the evaluation metrics used, with a call for greater focus on out-of-domain generalization and ethically-aligned evaluation.

The future of the field is bright but challenging. The path forward leads into the complex territories of linguistic nuance, such as sarcasm and irony; into the "black box" of the models themselves, demanding greater transparency and interpretability; and outward into a more global and holistic view of emotion, embracing multilingual and multimodal approaches. The journey of affective computing continues, moving steadily towards its ultimate goal: a more complete, human-centric, and truly intelligent understanding of emotion in all its forms.

#### Works cited

1. Detection of emotion by text analysis using machine ... - Frontiers, accessed June 22, 2025, <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1190326/full>
2. Detection of emotion by text analysis using machine learning - PMC, accessed June 22, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10548207/>
3. Emotion Analysis of Social Media Data using Machine Learning Techniques - IOSR Journal, accessed June 22, 2025, <https://www.iosrjournals.org/iosr-jce/papers/Conf.17015-2017/Volume-2/4.%2017-20.pdf>
4. Naive Bayes vs. SVM for Text Classification - GeeksforGeeks, accessed June 22, 2025, <https://www.geeksforgeeks.org/machine-learning/naive-bayes-vs-svm-for-text-classification/>
5. Mastering NLP Models for Text Classification - MyScale, accessed June 22, 2025, <https://myscale.com/blog/nlp-models-text-classification-insights/>
6. emotion detection from tweets using a bert and ... - VAPOR LIQUID, accessed June 22, 2025, <https://www.scientificbulletin.upb.ro/rev_docs_arhiva/fullf01_929427.pdf>
7. Feature Engineering Explained | Built In, accessed June 22, 2025, <https://builtin.com/articles/feature-engineering>
8. Enhancing Text Classification Through Grammar-Based Feature Engineering and Learning Models - MDPI, accessed June 22, 2025, <https://www.mdpi.com/2078-2489/16/6/424>
9. Feature Engineering for Linguistic Analysis - Number Analytics, accessed June 22, 2025, <https://www.numberanalytics.com/blog/feature-engineering-for-linguistic-analysis>
10. Feature Engineering – Text Data | Foundations of AI & ML, accessed June 22, 2025, <https://mylearningsinaiml.wordpress.com/ml-models/feature-engineering/feature-engineering-text-data/>
11. Sentiment Analysis using Transformer Models: Understanding Emotion in Text, accessed June 22, 2025, <https://jaimlr.github.io/Journal-of-Artificial-Intelligence-Machine-Learning-Research/sentiment-analysis-using-transformer-models-by-loveleen-narang.html>
12. Emotion Detection with Transformers: A Comparative Study - arXiv, accessed June 22, 2025, <https://arxiv.org/pdf/2403.15454>
13. Transformers Language Interpretability | AI Lab - The Smart Cube, accessed June 22, 2025, <https://www.thesmartcube.com/ai-lab/experiments/transformers-language-interpretability/>
14. (PDF) Emotion Detection using Transformer Model with Deep Learning - ResearchGate, accessed June 22, 2025, <https://www.researchgate.net/publication/390266831_Emotion_Detection_using_Transformer_Model_with_Deep_Learning>
15. (PDF) Transformer Models for Text-based Emotion Detection: A ..., accessed June 22, 2025, <https://www.researchgate.net/publication/348740926_Transformer_Models_for_Text-based_Emotion_Detection_A_Review_of_BERT-based_Approaches>
16. BERT and RoBERTa Based Sentiment Analysis - CorestratAI | AI SaaS, accessed June 22, 2025, <https://www.corestratai.com/post/comparative-study-on-bert-and-roberta-based-sentiment-analysis>
17. View of Comparative Analysis of Transformers to Support Fine ..., accessed June 22, 2025, <https://journals.flvc.org/FLAIRS/article/view/130612/133913>
18. MLP Architectures for Emotion and Sentiment Analysis: A Comparative Study of Fine-Tuned BERT and RoBERTa Models - OpenReview, accessed June 22, 2025, <https://openreview.net/pdf/d4befaf2facd2c347aa15cc0d07baa2639ac2747.pdf>
19. Overview of ROBERTa model - GeeksforGeeks, accessed June 22, 2025, <https://www.geeksforgeeks.org/machine-learning/overview-of-roberta-model/>
20. Emotion Classification In Software Engineering Texts: A ... - arXiv, accessed June 22, 2025, <https://arxiv.org/pdf/2401.10845>
21. Emotion Detection in Text: A Comprehensive Analysis Using Classical, Deep Learning, and Transformer-Based Models - NORMA@NCI Library, accessed June 22, 2025, <https://norma.ncirl.ie/7963/>
22. Comparative Analysis of Machine Learning Models for Emotion Classification in Textual Data | Airlangga | Kesatria : Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen) - tunasbangsa.ac.id, accessed June 22, 2025, <https://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/383>
23. Analysing Social Media Data for Emotion Detection and Sentiment Analysis - ResearchGate, accessed June 22, 2025, <https://www.researchgate.net/publication/386089533_Analysing_Social_Media_Data_for_Emotion_Detection_and_Sentiment_Analysis>
24. (PDF) A Comparative Analysis of MultinomialNB, SVM, and BERT ..., accessed June 22, 2025, <https://www.researchgate.net/publication/386171567_A_Comparative_Analysis_of_MultinomialNB_SVM_and_BERT_on_Garuda_Indonesia_Twitter_Sentiment>
25. Emotion Detection with Pre-Trained Language Models BERT and ..., accessed June 22, 2025, <https://imiens.org/index.php/imiens/article/view/45>
26. Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA - arXiv, accessed June 22, 2025, <https://arxiv.org/pdf/2104.02041>
27. Comparative Analysis of Transformer Models for Multi-Class Text ..., accessed June 22, 2025, <https://norma.ncirl.ie/7921/1/sharanyaneelakanti.pdf>
28. Emotion Detection in Reddit: Comparative Study of Machine Learning and Deep Learning Techniques - arXiv, accessed June 22, 2025, [https://arxiv.org/pdf/2411.10328?](https://arxiv.org/pdf/2411.10328)
29. Emotion Detection in Reddit: Comparative Study of Machine ... - arXiv, accessed June 22, 2025, <https://arxiv.org/pdf/2411.10328>
30. View of Emotion Classification from Covid-19Pandemic Tweets ..., accessed June 22, 2025, <https://jisem-journal.com/index.php/journal/article/view/5336/2502>
31. [2411.10328] Emotion Detection in Reddit: Comparative Study of Machine Learning and Deep Learning Techniques - arXiv, accessed June 22, 2025, <https://arxiv.org/abs/2411.10328>
32. A Review of Sentiment Analysis and Emotion Detection from Text using Different Models, accessed June 22, 2025, <https://www.researchgate.net/publication/387856791_A_Review_of_Sentiment_Analysis_and_Emotion_Detection_from_Text_using_Different_Models>
33. (PDF) Applying Transformer Models for Classifying Emotions in ..., accessed June 22, 2025, <https://www.researchgate.net/publication/390635825_Applying_Transformer_Models_for_Classifying_Emotions_in_Social_Media_Content>
34. Text Augmentation-Based Model for Emotion Recognition Using Transformers, accessed June 22, 2025, <https://www.techscience.com/cmc/v76n3/54359/html>
35. Comparative Analysis of Transformers to Support Fine-Grained Emotion Detection in Short-Text Data - ResearchGate, accessed June 22, 2025, <https://www.researchgate.net/publication/360382707_Comparative_Analysis_of_Transformers_to_Support_Fine-Grained_Emotion_Detection_in_Short-Text_Data>
36. Learning More from Mixed Emotions: A Label Refinement Method for Emotion Recognition in Conversations | Transactions of the Association for Computational Linguistics - MIT Press Direct, accessed June 22, 2025, <https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00614/118715/Learning-More-from-Mixed-Emotions-A-Label>
37. Challenges and Opportunities of Text-Based Emotion Detection: A Survey - ResearchGate, accessed June 22, 2025, <https://www.researchgate.net/publication/377543172_Challenges_and_Opportunities_of_Text-based_Emotion_Detection_A_Survey>
38. 12 dataset results for Emotion Recognition in Conversation AND Texts - Papers With Code, accessed June 22, 2025, <https://paperswithcode.com/datasets?task=emotion-recognition-in-conversation&mod=texts&page=1>
39. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification - ACL Anthology, accessed June 22, 2025, <https://aclanthology.org/2020.findings-emnlp.148.pdf>
40. Applying and Understanding an Advanced, Novel Deep Learning Approach: A Covid 19, Text Based, Emotions Analysis Study - PMC, accessed June 22, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8225489/>
41. SAMSEMO: New dataset for multilingual and ... - ISCA Archive, accessed June 22, 2025, <https://www.isca-archive.org/interspeech_2024/bujnowski24_interspeech.pdf>
42. Topic and Style-aware Transformer for Multimodal Emotion Recognition - ACL Anthology, accessed June 22, 2025, <https://aclanthology.org/2023.findings-acl.130.pdf>
43. Top 7 Metrics to Evaluate Sentiment Analysis Models - Focal, accessed June 22, 2025, <https://www.getfocal.co/post/top-7-metrics-to-evaluate-sentiment-analysis-models>
44. mteb/emotion · Datasets at Hugging Face, accessed June 22, 2025, <https://huggingface.co/datasets/mteb/emotion>
45. emotion Benchmark (Text Classification) - Papers With Code, accessed June 22, 2025, <https://paperswithcode.com/sota/text-classification-on-emotion>
46. Sarcasm Detection in Text: A Case Study Using Transformer Models, accessed June 22, 2025, <https://hunterheidenreich.com/posts/sarcasm-detection-with-transformers/>
47. Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores - MDPI, accessed June 22, 2025, <https://www.mdpi.com/2076-3417/14/21/9863>
48. Sentiment Analysis Metrics: Key Considerations - Insight7 - AI Tool For Interview Analysis & Market Research, accessed June 22, 2025, <https://insight7.io/sentiment-analysis-metrics-key-considerations/>
49. Emotion detection of social data: APIs comparative study - PMC, accessed June 22, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10172785/>
50. Emotion Detection in Social Media - Number Analytics, accessed June 22, 2025, <https://www.numberanalytics.com/blog/emotion-detection-in-social-media>
51. Unpacking Sarcasm: A Contextual and Transformer-Based ... - MDPI, accessed June 22, 2025, <https://www.mdpi.com/2073-431X/14/3/95>
52. A Transformer and Prototype-based Interpretable Model for Contextual Sarcasm Detection, accessed June 22, 2025, <https://arxiv.org/html/2503.11838v1>
53. Sarcasm Unveiled: A Comprehensive Systematic Review of Transformer-Based Detection Models, accessed June 22, 2025, <https://journal.esrgroups.org/jes/article/download/6679/4630/12312>
54. (PDF) Improving Attention-Based Interpretability of Text ..., accessed June 22, 2025, <https://www.researchgate.net/publication/363765786_Improving_Attention-Based_Interpretability_of_Text_Classification_Transformers>
55. Linguistic Interpretability of Transformer-based Language Models: a systematic review, accessed June 22, 2025, <https://arxiv.org/html/2504.08001v1>
56. Contrast-CAT: Contrasting Activations for Enhanced Interpretability in Transformer-based Text Classifiers | OpenReview, accessed June 22, 2025, <https://openreview.net/forum?id=vaQMMxhCqy>
57. Sentiment Analysis and Emotion Detection Using Transformer Models in Multilingual Social Media Data - The Science and Information (SAI) Organization, accessed June 22, 2025, <https://thesai.org/Downloads/Volume16No3/Paper_32-Sentiment_Analysis_and_Emotion_Detection.pdf>
58. Evaluating Arabic Emotion Recognition Task Using ChatGPT Models: A Comparative Analysis between Emotional Stimuli Prompt, Fine-Tuning, and In-Context Learning - MDPI, accessed June 22, 2025, <https://www.mdpi.com/0718-1876/19/2/58>
59. Survey on Current Trend in Emotion Recognition Techniques Using Deep Learning, accessed June 22, 2025, <https://www.researchgate.net/publication/370956908_Survey_on_Current_Trend_in_Emotion_Recognition_Techniques_Using_Deep_Learning>
60. "Only ChatGPT gets me": An Empirical Analysis of GPT versus other Large Language Models for Emotion Detection in Text - arXiv, accessed June 22, 2025, <https://arxiv.org/pdf/2503.04831>