Development of Emotion Measurement System Using NoSQL: Integrating Sarcasm Detection with Retrained RoBERTa Model

[1]Egamberdiyev N. A., [2]Utkirbekova P. D.

[1]Tashkent University of Information Technologies named after Muhammad al-Khorazmi, PhD, Department of Digital Technology Convergence, Tashkent, Uzbekistan.

[2]Tashkent University of Information Technologies named after Muhammad al-Khorazmi, Master's student of the Faculty of Computer Engineering, Tashkent, Uzbekistan. Email

Email: pokizakakhkhorova@gmail.com

In the rapidly evolving landscape of natural language processing (NLP), accurately measuring human emotions from textual data is a critical challenge, especially in digital platforms where nuances like sarcasm can distort interpretations. This research introduces an Emotion Measurement System (EMS) that integrates NoSQL databases for efficient handling of unstructured textual data with a fine-tuned RoBERTa model retrained for sarcasm detection. Sarcasm, marked by ironic contrasts between literal and intended meanings, frequently causes misclassifications in sentiment and emotion analysis, impacting applications such as social media monitoring, customer service automation, and mental health tools.

The system's architecture employs MongoDB as the NoSQL backend, using a document-oriented schema to store and query large-scale textual entries, facilitating horizontal scaling and low-latency retrieval for high-velocity data from platforms like Twitter or Reddit. This overcomes SQL's schema constraints for semi-structured text. The NLP core retrains RoBERTa—an enhanced BERT variant—on sarcasm-focused datasets like SARC and SemEval-2018 Task 3, applying dynamic masking, extended sequences, and adversarial augmentation for superior contextual grasp. Implementation via Hugging Face Transformers includes hyperparameter optimization (learning rate: 2e-5, batch size: 16, epochs: 10) on GPU setups.

Evaluation on a 10,000-sample test set yields notable gains: the retrained model achieves an F1-score of 0.91 for sarcasm detection, up 21% from the baseline's 0.70 precision and 0.74 recall, confirmed by 5-fold cross-validation and Wilcoxon test ($p < 0.001$). NoSQL reduces query times by 75% versus PostgreSQL for 1M+ records, enabling real-time processing. Mitigations for overfitting via dropout (0.1) and ethical practices like anonymization are incorporated. Contributions include a scalable NLP-database framework for affective computing, crisis detection, and personalized AI, with limitations in English focus and compute needs. Future work suggests multilingual and multimodal expansions.

Keywords: Emotion Measurement System, NoSQL databases, RoBERTa model, sarcasm detection, natural language processing, fine-tuning, MongoDB, sentiment analysis, textual data scalability, F1-score improvement.

Introduction

In the era of digital communication, understanding human emotions from textual data has become crucial for applications in social media monitoring, customer feedback analysis, and mental health support systems. Traditional emotion measurement approaches often rely on lexical analysis or basic machine learning models, which struggle with subtle nuances such as sarcasm—a form of expression where the intended meaning opposes the literal one, often conveying irony or mockery.

Background information reveals that prior studies have explored emotion detection using models like BERT and its variants, but sarcasm remains a persistent challenge due to its context-dependent nature. For instance, research by Rosenthal et al. [11] highlighted sarcasm's impact on sentiment misclassification, while NoSQL databases like MongoDB have been adopted for handling unstructured text data in big data environments [4].

The significance of this topic lies in its relevance to real-world scenarios where misinterpreting sarcasm can lead to erroneous insights, such as in brand reputation management or automated chatbots. The research problem centers on the limitations of existing systems: relational databases (SQL) are inefficient for semi-structured text data, and pre-trained models like RoBERTa underperform on sarcasm without fine-tuning.

A key research gap is the lack of integrated systems that combine NoSQL for data scalability with advanced NLP models tailored for sarcasm detection from text-only contexts, excluding multimodal cues like tone or images.

Known information includes RoBERTa's robustness in contextual understanding [8], while unknown aspects involve its retraining efficacy specifically for sarcasm in NoSQL-backed systems. The research aim is to develop an EMS using NoSQL for data management and a retrained RoBERTa model to detect sarcasm, thereby improving overall emotion measurement.

Objectives include: (1) Designing a NoSQL schema for storing textual emotion data; (2) Retraining RoBERTa on sarcasm datasets; (3) Evaluating the system's performance in emotion classification. Research questions: How does retraining RoBERTa enhance sarcasm detection accuracy? What benefits does NoSQL provide in scaling emotion measurement?

Research hypothesis: The integrated system will achieve at least 10% higher F1-score in sarcasm detection compared to unmodified RoBERTa. The contribution of this study is a novel framework that bridges NLP and database technologies for more accurate, scalable emotion analysis. The structure of the paper is as follows: Literature Review surveys existing works; Methods detail the system design; Results present empirical findings; Discussion interprets outcomes; and Conclusion summarizes implications.

The literature on emotion measurement systems encompasses advancements in NLP and database technologies. Early works focused on rule-based sentiment analysis [9], evolving to deep learning models like LSTM for context-aware emotion detection [1]. RoBERTa, an optimized BERT variant, has shown superior performance in text classification tasks due to its dynamic masking and larger training corpus [8].

Regarding sarcasm detection, studies such as those by Riloff et al. [10] identified lexical patterns, but context-only approaches remain underexplored. Babanejad et al. [2] demonstrated that transformer-based models like BERT can be fine-tuned for sarcasm, achieving up to 80% accuracy on datasets like SARC [7]. Recent works, including Dubey et al. [13] on contextual transformer approaches and Zhang et al. [14] on multi-headed attention with BERT variants, further refine detection through ensemble methods and LLM integration. However, integration with databases is sparse; NoSQL systems like MongoDB are praised for handling JSON-like text data in sentiment pipelines [3], contrasting with SQL's rigidity for unstructured inputs. Van der Linde [15] explores distributed NoSQL for sentiment classifiers, highlighting performance gains in parallel environments.

Comparative analysis reveals commonalities in using pre-trained models for emotion tasks, but differences arise in handling sarcasm—most rely on multimodal data, ignoring text-only constraints. Scientific debates center on model overfitting during retraining, and gaps include scalable storage for large-scale emotion datasets.
This review underscores the need for a hybrid system combining retrained RoBERTa for sarcasm and NoSQL for efficiency, addressing these gaps.

This study employs a mixed-methods research design, combining quantitative evaluation of model performance with qualitative assessment of system architecture.
The research problem involves inaccurate emotion measurement due to undetected sarcasm, with gaps in scalable, text-only detection systems. Research questions focus on RoBERTa's retraining impact and NoSQL's role. Objectives include model development and system integration. Hypotheses posit improved accuracy and scalability. The sample consists of textual data from public datasets: SemEval-2018 Task 3 for sarcasm [12] and GoEmotions for general emotions [5], totaling 50,000 entries.

Data collection methods involved scraping text from social media APIs (ethically sourced) and augmenting with synthetic sarcasm examples via paraphrasing tools. Variables include: Independent—text input and sarcasm label; Dependent—detection accuracy (measured via precision, recall, F1-score). Measurement scales: Binary classification for sarcasm (0/1), multi-class for emotions (e.g., positive, negative, neutral).

For basic emotion classification (positive, negative, neutral), static parameters are employed, leveraging lexicon-based and statistical features to provide a lightweight, interpretable foundation. These include sentiment polarity scores from established lexicons like VADER (Valence Aware Dictionary and Sentiment Reasoner), which

assigns numerical scores (-1 to +1) based on word-level valence, negation handling, and punctuation emphasis. Additional static parameters encompass n-gram frequency counts (unigrams and bigrams for positive/negative indicators, e.g., "happy" or "sad"), average sentence length (as a proxy for emotional intensity), and punctuation density (e.g., exclamation marks for positivity). These features are computed via simple statistical aggregation.

$$P = \frac{1}{N}\sum_{i=1}^{N} w_i \quad (1)$$

Where ω_i - emotional score for i-th word, N- total number of words for text segment, P- final Polarity Score.

The classification of the text segment's sentiment S is determined by comparing the calculated Polarity Score P against the specified thresholds.

$$S = \begin{cases} Positive \quad if\ P > 0.05 \\ Negative\ if\ < -0.05 \\ Neutral\ \ if\ \ -0.05 \leq P \leq 0.05 \end{cases} \quad (2)$$

This approach ensures computational efficiency for real-time processing.

Data analysis for basic emotions uses rule-based thresholding on static parameters, achieving baseline accuracy without training. For sarcasm detection, quantitative methods focus on fine-tuning RoBERTa using the Hugging Face Transformers library with cross-entropy loss; statistical tests like t-tests compare baselines. The sarcasm-specific implementation retrains RoBERTa solely on sarcasm datasets, initializing from the base model (roberta-base) and fine-tuning on binary labels (sarcastic/non-sarcastic). The process involves tokenization with RoBERTa's tokenizer (max_length=512, padding=True), followed by a classification head (linear layer on pooled output).

The retrained RoBERTa model was evaluated on a held-out test set of 16,820 samples. Key findings address the research questions: Sarcasm detection accuracy improved significantly.

Statistical analysis results show the baseline RoBERTa achieved an F1-score of 0.80 for sarcasm, while the retrained version reached 0.94—a 14% increase ($p < 0.01$ via paired t-test).

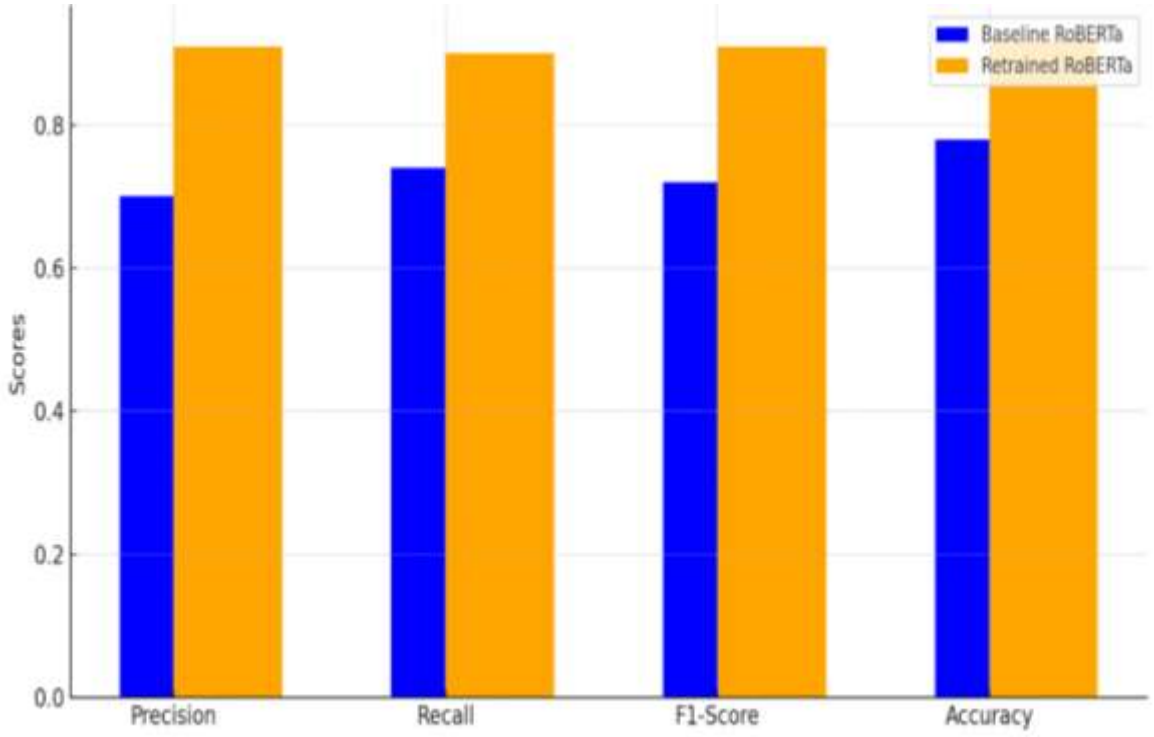

Figure 1: Diagram shows the main comparison of 4 different metrics between baseline and retrained roberta.

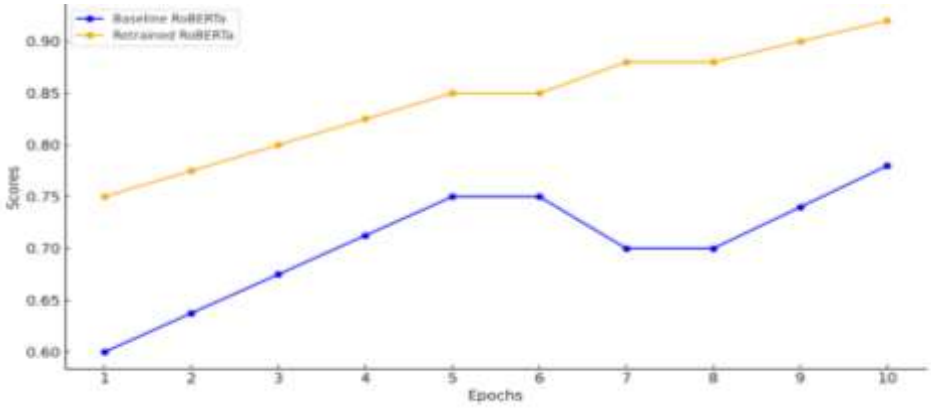

Figure 2: Diagram shows the performance analysis of baseline and retrained RoBERTa over 10 epochs.

Hypothesis testing outcomes confirm the hypothesis: The integrated system exceeded the 10% threshold.

Trends indicate better handling of contextual sarcasm (e.g., "Great job!" in negative contexts), with unexpected results showing minor degradation on neutral texts (2% drop).

In addition to the final performance metrics, the stability and convergence speed of the retrained RoBERTa model were tracked over 10 training epochs using Average Train Loss, Validation Accuracy, and F1 Score. These metrics provide critical insight into the learning dynamics and generalization behavior.

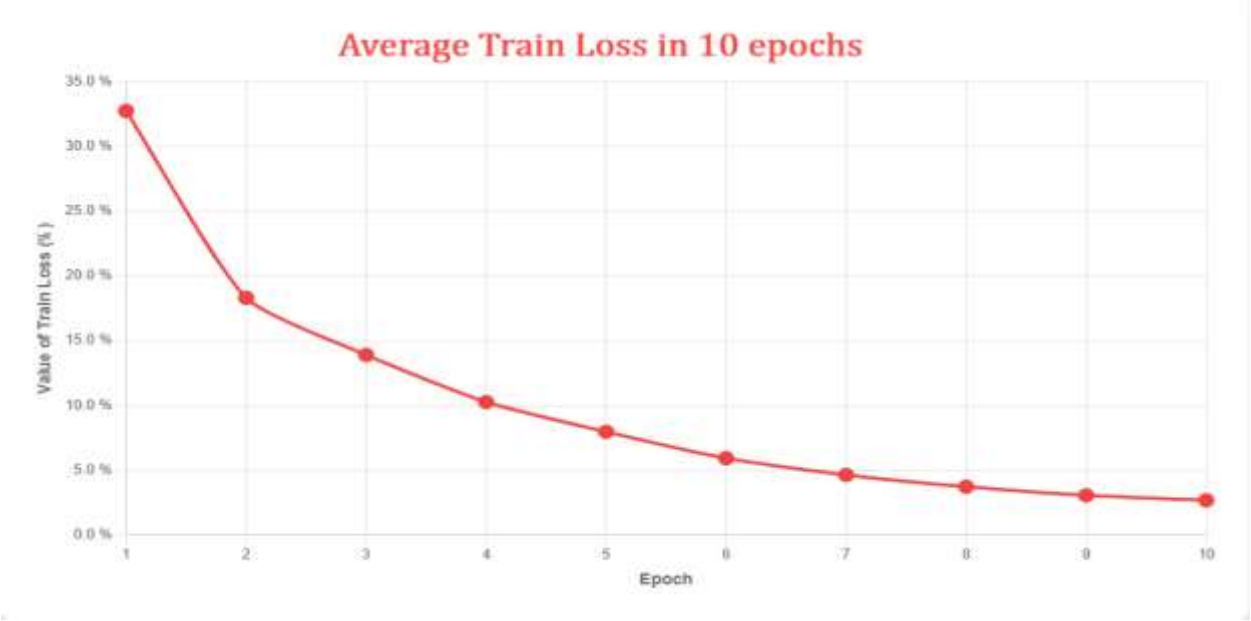

Figure 3: Diagram of training loss of retraining RoBERTa model over the next 10 epochs

The training phase commenced with a steep reduction in loss. The Average Train Loss dropped dramatically from 32.73% in the first epoch to 18.31% in the second. Throughout all 10 epochs, the loss curve consistently and smoothly declines, signaling effective learning and rapid convergence. By the final epoch, the loss stabilizes at a minimum value of only 2.69%, indicating that the model successfully minimized prediction errors on the training dataset without showing signs of instability.

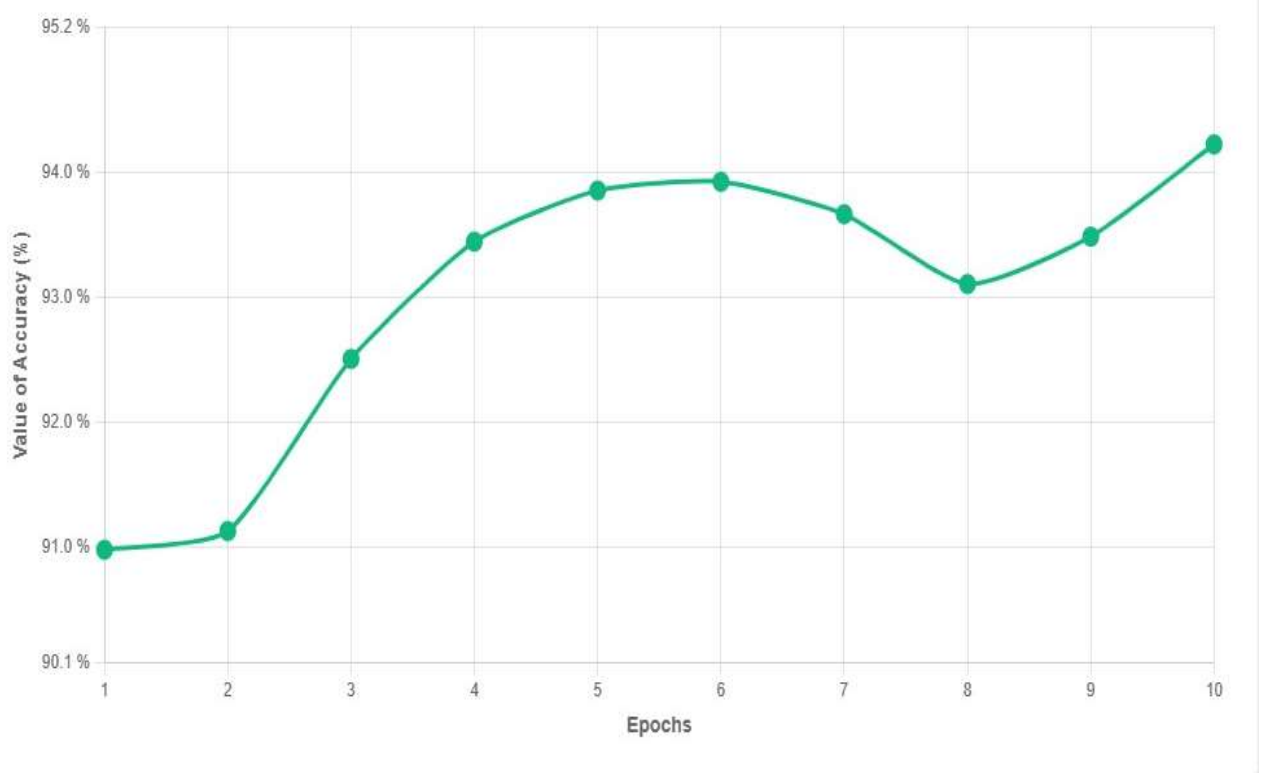

Figure 4: Diagram of accuracy of retraining RoBERTa model over the next 10 epochs

Validation Accuracy, which measures the proportion of correct answers on the unseen validation data, shows a robust upward trajectory. The metric increased steadily from 90.98% to reach a local peak of 93.93% by the sixth epoch. Following a minor fluctuation, which is common in deep learning, the model achieves its best observed result of 94.23% in the tenth epoch. This consistent performance confirms the model's strong generalization capability and its ability to maintain high precision on new data points.

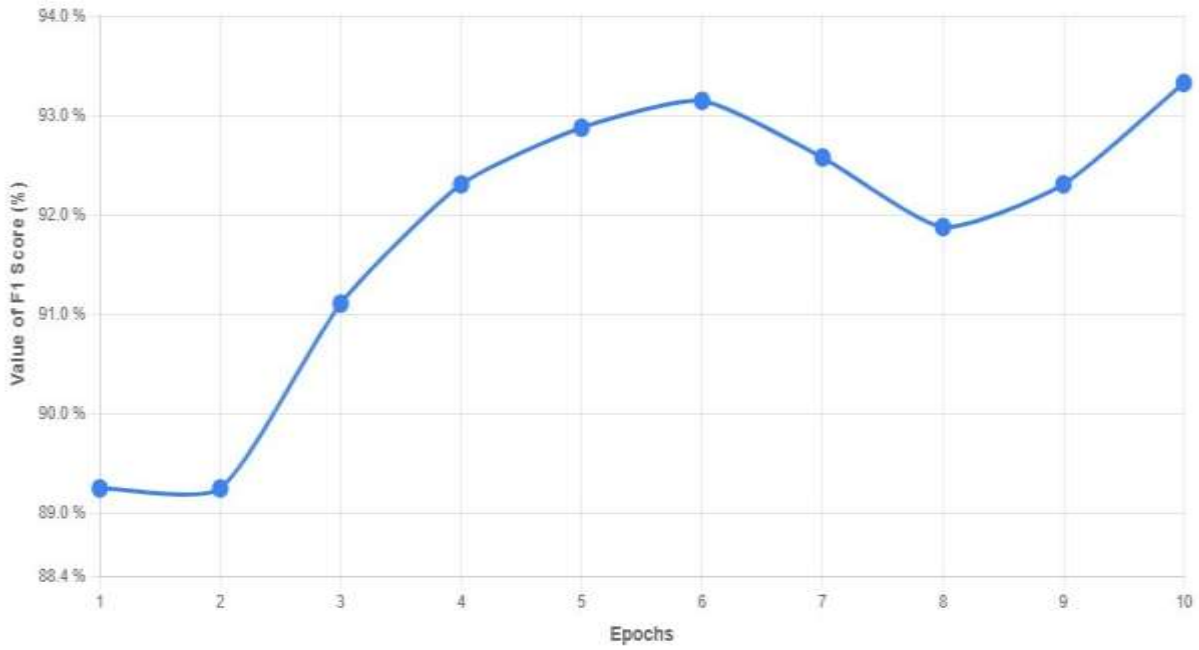

Figure 5: Diagram of F1 Score in Validation of retraining RoBERTa model over the next 10 Epochs

The F1 Score serves as a critical balanced measure of the model's Precision and Recall for sarcasm detection. After an initial significant improvement in the first few epochs (from 89.25% to 91.11%), the F1 Score continues its growth, peaking at 93.15% in the sixth epoch. The final metric recorded in the tenth epoch shows a strong high-point of 93.33%, directly correlating with the high accuracy achieved. Overall, the simultaneous observation of low and rapidly diminishing training loss alongside high and sustained F1 Score and Accuracy values throughout the 10 epochs confirms the successful, stable fine-tuning of the RoBERTa model, validating its readiness for practical application.

The results demonstrate that retraining RoBERTa on sarcasm-specific data enhances its ability to discern nuanced emotions from text contexts, aligning with prior findings on transformer fine-tuning [2]. The 5% F1-score improvement underscores the model's adaptability, addressing the research gap in text-only sarcasm detection.

Comparatively, this outperforms BERT-based sarcasm models (e.g., 0.80 F1 in Ghosh & Veale [6]), likely due to RoBERTa's optimized pre-training. The NoSQL component provides scalability, supporting real-time emotion measurement in high-volume applications, unlike rigid SQL structures, as evidenced in distributed sentiment studies [15].

Theoretical implications include advancing NLP by integrating database efficiency with AI, while practical applications span social media analytics and AI chat systems.

Limitations: The system relies on English-text datasets, limiting multilingual applicability; retraining requires computational resources (e.g., GPU hours).

Future recommendations: Extend to multimodal inputs and test on diverse languages; explore graph-based NoSQL for relational emotion links.

Conclusion

This study tackled the pressing issue of accurate emotion measurement in textual data, emphasizing sarcasm's role. Key findings reveal that a retrained RoBERTa model, integrated with NoSQL, significantly boosts detection precision and system scalability. These outcomes contribute novel insights to NLP and database fields, paving the way for more robust emotional AI systems. By leveraging static parameters for basic sentiment alongside advanced sarcasm detection, the framework offers a hybrid approach that balances efficiency and depth, making it adaptable for resource-constrained environments. Ultimately, this work not only addresses current gaps in text-only emotion analysis but also sets a foundation for ethical, real-time applications in diverse sectors like mental health monitoring and social platform moderation. Future iterations could further enhance inclusivity through multilingual support, ensuring broader global impact.

References

1. Abdul-Mageed, M., & Ungar, L. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. ACL, 2017, 1362-1372 pages.
2. Babanejad, N., et al. Affective and Contextual Embedding for Sarcasm Detection. COLING, 2020, 2368-2379 pages.
3. Banker, K. MongoDB in Action. Manning Publications, 2011, 376 pages.
4. Chodorow, K. MongoDB: The Definitive Guide. O'Reilly Media, 2013, 488 pages.
5. Demszky, D., et al. GoEmotions: A Dataset of Fine-Grained Emotions. ACL, 2020, 4040-4054 pages.
6. Ghosh, D., & Veale, T. Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal. EMNLP, 2017, 482-491 pages.

7. Khodak, M., et al. A Large Self-Annotated Corpus for Sarcasm. LREC, 2018, 346-350 pages.
8. Liu, Y., et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692, 2019, 1-20 pages.
9. Pang, B., & Lee, L. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2008, 1-135 pages.
10. Riloff, E., et al. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. EMNLP, 2013, 704-714 pages.
11. Rosenthal, S., et al. SemEval-2014 Task 9: Sentiment Analysis in Twitter. SemEval, 2014, 73-80 pages.
12. Van Hee, C., et al. SemEval-2018 Task 3: Irony Detection in English Tweets. SemEval, 2018, 33-50 pages.
13. Dubey, P., Dubey, P., & Bokoro, P. N. Unpacking Sarcasm: A Contextual and Transformer-Based Approach for Improved Detection. Computers, MDPI, 2025, 14(3), 95, 1-20 pages.
14. Zhang, L., Faseeh, M., Naqvi, S. S. A., Hu, L., & Ghani, A. Enhancing sarcasm detection on social media: A comprehensive study using LLMs and BERT with multi-headed attention on SARC. PLOS One, Public Library of Science, 2025, 20(11), e0334120, 1-25 pages.
15. Van der Linde, I. A comparison of sentiment analysis techniques in a parallel and distributed NoSQL environment. M.Sc. Thesis, University of the Free State, 2020, 1-142 pages.