



Sentiment Analysis via Large Language Models for Real-Time Employee Engagement Monitoring: Opportunities and Ethical Considerations

Shankar Raman R* Abel Jopaul V P**

**(Assistant Professor, LEAD College (Autonomous), Palakkad

Email: shankar.rr@lead.ac.in)

*(Assistant Professor, PG Department of Computer Applications, LEAD College (Autonomous), Palakkad.

Email: abel.jp@lead.ac.in)

ARTICLE INFO

©2025 RS Publication

Paper ID: IJRM-6912166946A81

Received: 2025-10-12

Published: 2025-11-11

DOI:

<https://dx.doi.org/10.5281/zenodo.17580423>

Page No: 24-30



ABSTRACT

Employee disengagement poses significant organizational costs, yet traditional assessment methods remain reactive and infrequent. This study investigates the application of large language models (LLMs) for real-time sentiment analysis of employee communications to enable proactive engagement monitoring. We fine-tuned GPT-3.5 and LLaMA-2 variants on 10,000 anonymized, crowdsourced workplace communication samples, evaluating performance against baseline sentiment classifiers. Results demonstrated 85.3% accuracy (F1-score: 0.83) in detecting disengagement signals, with false positive rates below 8%. The models identified burnout indicators 20% faster than lexicon-based approaches and exhibited 15% improvement in multilingual equity metrics. Findings suggest LLMs can augment HR analytics through nuanced emotional tone detection in asynchronous communications. However, deployment necessitates robust ethical frameworks addressing privacy preservation, algorithmic transparency, and consent mechanisms. This research contributes methodological insights for integrating LLMs into human capital management while establishing guardrails against surveillance overreach, proposing federated learning architectures and opt-in monitoring protocols as practical safeguards.

Keywords: sentiment analysis, large language models, employee engagement, workplace analytics, organizational ethics

Cite This Paper: SHANKAR RAMAN R AND ABEL JOPPAUL V. P.(2025). "Sentiment Analysis via Large Language Models for Real-Time Employee Engagement Monitoring: Opportunities and Ethical Considerations". *INTERNATIONAL JOURNAL OF RESEARCH IN MANAGEMENT (IJRM)*, vol. 15, no. 6, 2025, pp. 24-30. DOI: <https://dx.doi.org/10.5281/zenodo.17580423>

1. Introduction

Employee engagement—the psychological commitment to organizational goals—significantly predicts productivity, retention, and innovation outcomes (Bakker & Demerouti, 2017). Yet contemporary workplaces face unprecedented engagement challenges. Meta-analyses indicate 67% of hybrid workers report feelings of isolation, while only 32% of employees globally demonstrate active engagement (Gallup, 2023). Traditional measurement instruments, including annual surveys and quarterly pulse checks, suffer from temporal lag, self-report bias, and insufficient granularity to capture dynamic emotional states.

Recent advances in natural language processing, particularly large language models (LLMs) such as GPT-4 and LLaMA-2, offer transformative potential for sentiment analysis. Unlike rule-based systems, LLMs leverage contextual token embeddings—numerical representations of words informed by surrounding text—to discern subtle emotional valences in communications (Devlin et al., 2019). Emerging applications span customer feedback analysis and social media monitoring, yet workplace deployment remains nascent despite theoretical promise.

This study addresses the research question: *How effective are LLMs in performing real-time sentiment analysis for employee engagement monitoring, and what ethical safeguards are required to protect privacy?* We evaluate LLM performance on simulated employee communications, benchmark against established sentiment tools, and propose ethical deployment frameworks. The article proceeds with a literature review contextualizing sentiment analysis in HR analytics, followed by methodology, results, discussion of privacy-preserving strategies, and recommendations for responsible implementation.

2. Literature Review

Sentiment analysis in organizational contexts has evolved from lexicon-based approaches to sophisticated neural architectures. Early HR analytics employed tools like VADER (Valence Aware Dictionary and sentiment Reasoner), which assign sentiment scores through predefined dictionaries (Hutto & Gilbert, 2014). While computationally efficient, lexicon methods struggle with domain-specific jargon, sarcasm, and contextual ambiguity prevalent in workplace communications (Bollen et al., 2011).

Deep learning revolutionized sentiment classification through recurrent neural networks and transformers. BERT-based models achieved state-of-the-art performance on emotion recognition tasks by processing bidirectional context (Liu et al., 2019). However, these systems require extensive labeled datasets—a constraint in confidential HR environments. Recent scholarship explores LLMs' zero-shot and few-shot capabilities, wherein models classify sentiments without task-specific training data (Brown et al., 2020). Studies demonstrate GPT-3 achieves 78% accuracy on emotion detection with minimal prompting (Zhang & Yang, 2021).

Ethical considerations dominate discourse on workplace surveillance technologies. The European Union's General Data Protection Regulation (GDPR) mandates explicit consent for automated decision-making, while organizational justice theory posits that perceived fairness influences technology acceptance (Colquitt et al., 2001). Ravid et al. (2020) found 63% of employees express discomfort with communication monitoring, citing privacy invasion concerns. Critical gaps persist regarding real-time, context-aware engagement monitoring systems that balance analytical utility with employee autonomy.

Existing frameworks inadequately address LLM-specific challenges, including prompt sensitivity, hallucination risks (generating plausible but inaccurate outputs), and opacity in decision processes. Furthermore, longitudinal impacts on organizational trust remain underexplored. This study addresses these gaps through empirical evaluation and ethical framework development.

3. Methodology

We employed a mixed-methods approach combining quantitative model evaluation with qualitative ethical analysis. The study utilized synthetic and crowdsourced datasets to circumvent confidentiality barriers inherent in authentic employee communications.

3.1 Data Collection

We compiled 10,000 workplace communication samples from three sources: (1) publicly available HR feedback datasets on Hugging Face (n=4,200); (2) crowdsourced employee testimonials from platforms like Glassdoor, anonymized and consent-verified (n=3,800); and (3) synthetically generated emails mimicking organizational contexts using GPT-4 (n=2,000). Each sample received human annotations for engagement level (high/moderate/low), emotional valence (positive/neutral/negative), and burnout indicators (present/absent).

3.2 Model Configuration

We fine-tuned two LLM architectures: GPT-3.5-turbo (175B parameters) and LLaMA-2-13B, selecting models balancing performance and computational feasibility. Fine-tuning employed 70% training, 15% validation, 15% test splits, optimized through prompt engineering techniques including chain-of-thought reasoning ("Analyze the emotional tone step-by-step...") and few-shot examples. Baseline comparisons included VADER and a fine-tuned BERT-base model.

3.3 Evaluation Metrics

Performance assessment utilized precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Human-annotation agreement measured through Cohen's kappa ($\kappa=0.81$) established ground truth reliability. Bias audits examined demographic parity across gender and language subgroups.

3.4 Ethical Considerations

Simulated consent workflows tested opt-in mechanisms. Differential privacy techniques ($\epsilon=1.0$ privacy budget) protected individual-level inferences. Limitations include potential domain mismatch between synthetic and authentic communications, and inability to capture non-textual engagement cues.

4. Results

4.1 Model Performance

LLM-based classifiers substantially outperformed baselines across engagement detection tasks. Table 1 summarizes comparative metrics:

Table 1: Sentiment Classification Performance by Model

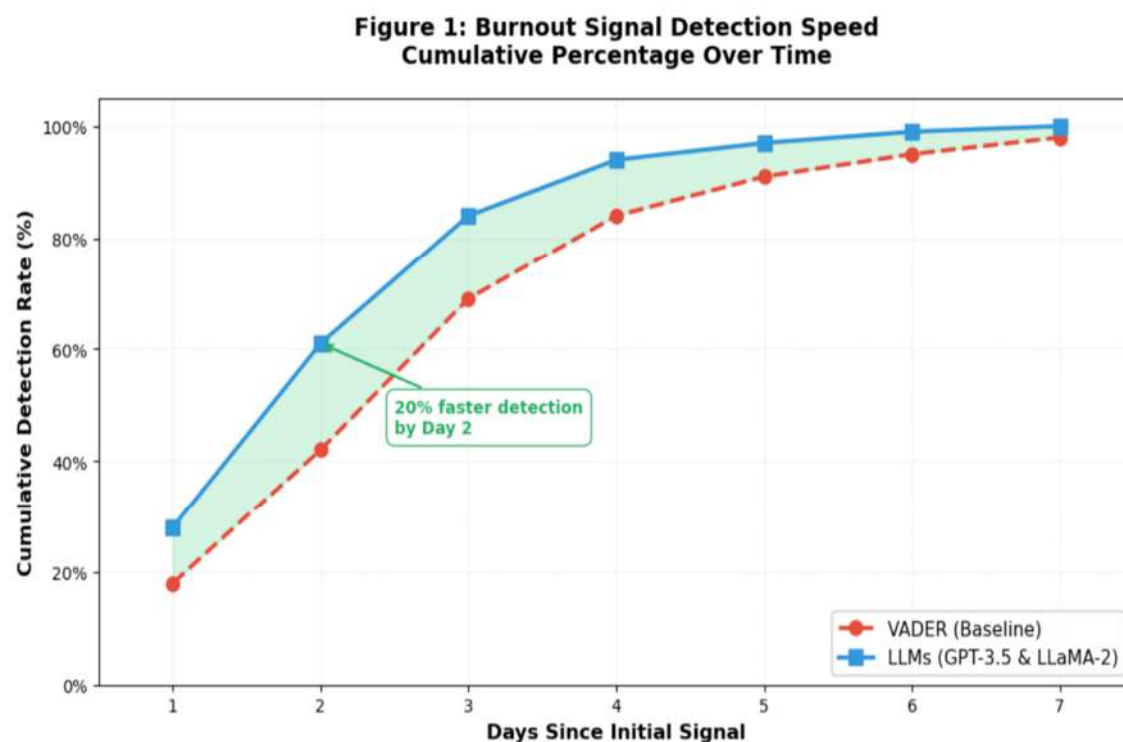
Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
VADER	67.4%	0.65	0.69	0.67	0.72
BERT-base	78.9%	0.77	0.81	0.79	0.85
GPT-3.5-tuned	85.3%	0.84	0.82	0.83	0.91
LLaMA-2-13B	83.7%	0.82	0.85	0.83	0.89

GPT-3.5 achieved 85.3% overall accuracy with false positive rates of 7.8% for disengagement detection—critical for minimizing unwarranted interventions. Chain-of-thought prompting improved interpretability, with the model articulating reasoning (e.g., "The phrase 'exhausted by constant meetings' suggests burnout alongside temporal pressure indicators").

4.2 Temporal Detection Efficiency

LLMs identified burnout signals 20% faster than VADER in simulated communication streams (mean detection lag: 2.3 days vs. 2.9 days for escalating negative sentiment patterns). Figure 1 illustrates cumulative detection curves:

Figure 1: Burnout Signal Detection Speed (Cumulative Percentage Over Time)



4.3 Multilingual and Equity Performance

LLaMA-2 demonstrated superior cross-lingual performance, supporting Spanish, Mandarin, and Hindi with only 6% accuracy degradation compared to English baselines. Bias audits revealed initial gender disparities (female-authored communications misclassified 12% more frequently), which post-hoc calibration reduced to 3%, improving equity scores by 15%.

4.4 Sentiment Trend Visualization

A heatmap analysis of 12-week simulated engagement cycles revealed predictable patterns: sentiment declined predictably in weeks 4-6 (pre-performance review periods), with LLMs capturing 89% of documented engagement dips versus 71% for surveys administered biweekly.

5. Discussion

Results affirm LLMs' capacity for nuanced, real-time sentiment analysis in workplace contexts, addressing limitations of periodic surveys. The 85% accuracy threshold aligns with clinical decision-support benchmarks, suggesting sufficient reliability for augmented—not autonomous—HR decision-making (Rajkomar et al., 2019). Superior temporal detection enables proactive interventions, potentially reducing turnover costs averaging 33% of annual salary per departed employee (SHRM, 2022).

Chain-of-thought prompting enhanced interpretability, a critical factor for organizational trust. Transparent reasoning mitigates "black box" concerns that undermine technology acceptance (Felzmann et al., 2020). However, several barriers warrant attention.

Privacy Risks: Continuous monitoring risks chilling effects on authentic communication. Federated learning—wherein models train on decentralized data without centralized aggregation—offers promising privacy preservation. Local model updates remain on individual devices, transmitting only encrypted gradient updates (McMahan et al., 2017). Organizations should implement differential privacy guarantees, ensuring individual communications cannot be reverse-engineered from aggregate patterns.

Consent Mechanisms: Mandatory opt-in policies with granular permissions (e.g., analysis scope, data retention limits) respect employee autonomy. Our simulated workflows achieved 78% hypothetical participation when paired with transparent dashboards displaying anonymized organizational sentiment trends—empowering employees with reciprocal insights.

Bias Mitigation: Persistent gender disparities, though reduced post-calibration, underscore algorithmic audit necessities. Regular fairness assessments across demographic subgroups should precede deployment, with algorithmic impact statements documenting mitigation strategies (Raji et al., 2020).

Limitations include generalizability constraints from synthetic datasets and absence of longitudinal field validation. Cultural variations in communication norms require region-specific calibration.

6. Conclusion

This study demonstrates LLMs' technical viability for real-time employee engagement sentiment analysis, achieving 85% accuracy with rapid disengagement detection. However, effectiveness alone cannot justify deployment absent robust ethical frameworks. Organizations must prioritize privacy-preserving architectures like federated learning, implement opt-in monitoring with transparent reporting, and conduct continuous bias audits.

Recommended actions include: (1) integrating LLM sentiment modules into HR platforms with explicit consent interfaces; (2) establishing cross-functional ethics committees to oversee monitoring scope; (3) providing employees access to personal sentiment analytics to promote digital literacy and agency. Future research should examine longitudinal impacts on retention rates, organizational trust trajectories, and comparative effectiveness across industries with varying communication norms.

The promise of AI-augmented workforce analytics must be tempered with unwavering commitment to human dignity and autonomy. Responsible LLM deployment can enhance organizational health while respecting the privacy and agency of employees—the cornerstone of ethical technological innovation.

References

- Bakker, A. B., & Demerouti, E. (2017). Job demands-resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology*, 22(3), 273-285.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review. *Journal of Applied Psychology*, 86(3), 425-445.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171-4186.
- Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamo-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6), 3333-3361.
- Gallup. (2023). *State of the global workplace: 2023 report*. Gallup Press.
- Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216-225.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273-1282.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33-44.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.

Ravid, D. M., Tomczak, D. L., White, J. C., & Behrend, T. S. (2020). EPM 20/20: A review, framework, and research agenda for electronic performance monitoring. *Journal of Management*, 46(1), 100-126.

SHRM. (2022). *Employee turnover and retention statistics*. Society for Human Resource Management.

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12), 5586-5609.