UARS: A Unified Employee Attrition Risk Controlling Model using Machine Learning and Transformer Based Sentiment Analysis

Sujal Das[1], Urbee Patra[2], Rehan Gupta[3], Sumit Paul[4], Takaaki Goto[5], Partha Ghosh*[6][0000-0003-2843-2668], Subhankar Roy[7][0000-0002-6307-7043], Soumya Sen[8][0000-0002-9178-6410],

[1,2,3,4,6,7]Academy of Technology, West Bengal, India
[5]Toyo University, Saitama, Japan
[8]University of Calcutta, Kolkata, India
{[1]dassujal83, [2]urbeepatra07, [3]guptarehan482, [4]paulsumit1011, [6]pghosh44, [7]subhankar.roy07, [8]iamsoumyasen}@gmail.com
[5]tg@gotolab.net

Abstract. Employee attrition poses a critical challenge for businesses, costing organizations significantly in recruitment, training, and lost productivity. Therefore, organizations must distinguish between valuable and non-valuable employees. Most importantly, organizations must prioritize employees, which depends upon several factors and is, in itself, a tedious task. In order to have a scientific employee retention policy, this paper introduces a three-phase model that significantly helps in proactive retention risk assessment by integrating advanced Machine Learning algorithms on structured human resources data with Natural Language Processing for sentiment analysis on unstructured employee feedback. In the first phase, the proposed model performs extensive data analysis to identify initial correlations among attributes that results employee attrition. We then train and evaluate several classification models, demonstrating robust predictive performance. Phase two establishes an Employee Value Assessment model, which combines objective performance metrics with attrition risk to classify employee importance, which later helps to simulate targeted compensation (salary hike) recommendations. In third phase, we integrate a Transformer model to derive objective emotional scores from simulated survey responses, providing deeper, qualitative insights into current employee morale. The combination of high-accuracy quantitative prediction along with real-time sentiment signals is formalized into a Unified Attrition Risk Score (UARS), offering a holistic and actionable approach to HR management, that enables targeted intervention strategies to improve retention and enhance employee well-being.

Keywords: Employee Attrition, Machine Learning, Deep Learning, Sentiment Analysis, BERT, Predictive Modelling, Feature Importance, XGBoost.

1 Introduction

Several recent studies show that employee attrition is one of the major problems in many reputed companies [1][9]. The cost of replacing an employee is estimated to be between half and twice the employee's annual salary. This depends on the employee’s seniority and specialization [1][3]. Financial impact is only one problem. But, beyond the direct financial impact, high turnover harms institutional knowledge, team morale and long-term productivity. Frequent employee resignations departures disrupt project continuity and increase the workload for remaining staff. This ultimately reduces organizational efficiency and innovation. Because of these effects, predicting and managing employee turnover has become an important part of modern Human Resource Management (HRM) [10][14]. Most traditional HR practices were to address the turnover only after resignations occurred. Although, most of the today's advanced organizations are moving toward proactive and data-driven strategies that predicts and tries to reduce risks before they become serious issues [4][6][7]. Despite these advances, many current approaches to manage employee turnover are often disconnected. Predictive models typically rely on historical structured data like age, position, and performance metrics. They often ignore the behavioural and emotional factors that can cause dissatisfaction [22]. Employee surveys can provide useful insights into sentiment. But they are infrequent and lack the complete picture. This gap shows the need for combined frameworks that merge machine learning with sentiment analysis [5][8][13][19]. Hence, employers can get ongoing, flexible and useful insights into employee retention. Therefore, the goal of this research is to address these gaps by creating an integrated analytical model. This model must leverage a robust quantitative model for risk scoring and an advanced qualitative component (like sentiment analysis) to diagnose the underlying causes. Finally, to generate a calculated Employee Value Metric to prioritize retention efforts strategically [7][9][11]. Therefore, the major goals of this research are:

a) To develop and evaluate several ML models to predict attrition risk on a representative HR dataset that further utilized for defining and implementing a quantifiable Employee Value Metric that cross-references predictive risk with an employee's organizational contribution.
b) To integrate a sentiment analysis based module to interpret open-ended employee feedback for real-time diagnostic insights.
c) Finally, to develop a prescriptive framework for targeted compensation and environmental improvements based on the combined quantitative and qualitative data streams.

The subsequent sections of this paper detail out the dataset used, step-by-step analytical methodologies, the integrated results and finally the prescriptive retention strategy derived from the fusion of these approaches.

2 Related Work

The study of employee turnover spans across decades: drawing heavily from fields such as organizational psychology, economics and more recently, data science. Early predictive models [1][7] in HR focused on organizational commitment and job satisfaction as primary drivers, typically using regression models. Introduction of Big data and massive computational resources [15] have caused a paradigm shift towards advanced machine learning. Studies frequently employ algorithms like Logistic Regression (due to its interpretability), Decision Trees or Random Forests (for handling non-linear relationships and feature interactions) and Gradient Boosting Machines (e.g., XGBoost) (for superior predictive performance) [14]. Key features for attrition [18] such as Monthly Income, Years At Company, Job Level, Environment Satisfaction and Over-Time are constantly identified in the literature [16][17]. Simply predicting who will leave is insufficient, therefore the organizations must know who they can least afford to lose. Therefore, Employee Value modelling [9] is required for this. Value models often combine factors like performance ratings, tenure, uniqueness of skill set, and replacement cost into a single index [10[11]. Compensation planning, traditionally based on market rates and annual reviews, is increasingly being optimized using predictive value metrics to ensure retention funds are spent efficiently [16]. The concept of linking a model's attrition risk output to a prescriptive action (like a salary hike) forms a crucial bridge between analytics and HR strategy. The development of Natural Language Processing (NLP) [8][12] provides a scalable, qualitative perspective on employee experience. Feedback from traditional employee surveys is frequently forced into similar scales, losing its individuality. Richer, real-time insights can be obtained by analysing open-ended text from internal communications and surveys. Recent work has utilized deep learning models like BERT (Bidirectional Encoder Representations from Transformers) [2][8] for highly accurate sentiment and emotion detection. It has been demonstrated that incorporating these sentiment scores straight into attrition models increases accuracy, especially when forecasting voluntary turnover caused by arbitrary elements like manager communication or work-life balance [6].

But all the previous models failed to provide the development and evaluation of several ML models for predicting attrition risk on a representative HR dataset that were further utilized for the definition and implementation of a quantifiable Employee Value metric that cross-references predictive risk with an employee's organizational contribution, the integration of a sentiment analysis-based module to interpret open-ended employee feedback for real-time diagnostic insights, and a prescriptive framework for targeted compensation and environmental improvements based on the combined quantitative and qualitative data streams. Therefore, the goal of this research is to integrate and establish quantitative ML techniques with modern NLP and a compensation simulation engine for creating a novel and holistic employee retention system.

3 Proposed Methodology

This section proposed a three phase employee retention policy. The figure-1 illustrates an integrated employee retention model combining ML-based attrition prediction, value scoring and transformer based sentiment analysis for salary hike decisions.

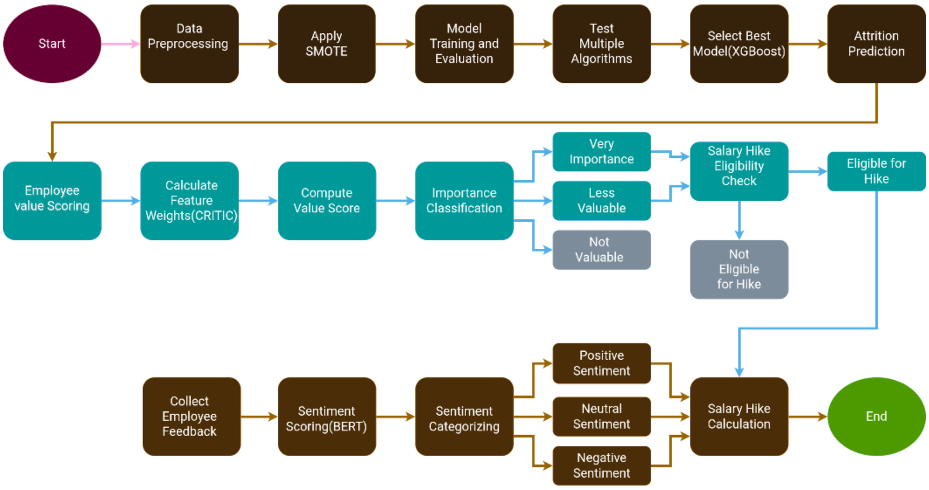

Figure-1. Flow diagram of the proposed model

Algorithm of the proposed model:

Purpose: Integration of Machine Learning (ML) prediction, Natural Language Processing (NLP) sentiment, and synthesis to identify high-risk, high-value employees and recommend targeted retention actions.

Input: Structured HR Data (e.g., Age, Monthly Income, Job Level), Unstructured Employee Feedback Text (e.g., internal survey responses).

Output: Prioritized list of at-risk employees and a specific, data-driven action plan.

Algorithm A: Attrition Risk Scorer (Predictive Model)

Input: Structured HR Data for one employee (e.g., age, monthly income, job level, department, etc.).

Output: Attrition Probability Score (**P_A** ranging from 0.0 to 1.0).

Step 1: Start

Step 2: Data Preparation

- **2.1.** Load employee HR records (e.g., Age, Monthly Income, Job Level).
- **2.2.** Drop non-predictive columns.
- **2.3.** Apply One-Hot Encoding to categorical features (e.g., Department, Job Role).
- **2.4.** Scale all numerical features using a Min-Max Scaler or Standard Scaler.

Step 3: Risk Prediction

- **3.1.** Load the pre-trained, optimized Machine Learning Model.
- **3.2.** Input the prepared feature vector for a specific employee into the model.
- **3.3.** The model computes the likelihood of Attrition.
- **3.4.** Output the Attrition Probability Score **P_A** (ranging from 0.0 to 1.0).

Step 4: End

Algorithm B: Value-Risk Action Generator

Input: Attrition Probability (**P_A** from Algorithm A), Structured HR Metrics (e.g., Job Level, Performance Rating, Total Working Years).

Output: Employee Value Category (**C_V**), Priority Index (**I_P**), and Recommended Retention Action (Intervention).

Step 1: Start

Step 2: Employee Value Calculation

2.1. Calculate a raw Employee Value Score **V_E** based on weighted HR metrics

2.2.C_V= Classify **V_E** into a categorical label with scores ranging from 1 to 10: 'Low Value', 'Medium Value', 'High Value'.

Step 3: Profile Synthesis

3.1. Combine two key inputs: Attrition Probability (**P_A** from algorithm A) and Employee Value (**C_V**).

3.2. Define a compound Priority Index **I_P**: **I_P** = f(**P_A**, **C_V**). (Highest priority for High **P_A**, and High **C_V**).

Step 4: End

Algorithm C: Sentiment Scoring (Transformer-Based)

Input: Unstructured Employee Feedback Text (multiple text responses from an employee's survey).

Output: Average Sentiment Score (ranging from -1.0 to +1.0) and Sentiment Category ('Negative', 'Neutral', 'Positive').

Step 1: Start

Step 2: Text Preprocessing

2.1. Give the employees a few questions and request that they respond to the survey.

2.2. Collect unstructured text responses from the employee's internal survey.

2.3. Clean raw text: remove punctuation, stop words, and normalize case.

Step 3: Sentiment Analysis

3.1. Apply a pre-trained Sentiment Analysis model (e.g., BERT-based model) to each individual survey response.

3.2. Record a Sentiment Score **S_I** for each response **i** (e.g., **-1.0** for negative, **+1.0** for positive).

Step 4: Aggregation and Classification

4.1. Calculate the employee's Average Sentiment Score **S_{avg}** by weighting and averaging the individual **S_I** scores.

4.2. Classify **S_{avg}** into a categorical label: 'Negative', 'Neutral' or 'Positive'.

4.3. Output the Average Sentiment Score **S_{avg}** and its category **C_S**.

Step 5: End

Integrated Workflow: The Retention Pipeline

Input: Structured HR Data (CSV/DB) and Unstructured Employee Feedback Text (DB/Survey) for all employees.

Output: Prioritized Action Plan (List of Employee IDs, their Priority Index I_P, and Recommended Action).

Step 1: Start

Step 2: Action Recommendation Engine

2.1. Apply a set of business rules based on the combined profile to recommend an action:

Rule 1: IF ($\mathbf{C_V}$ = 'High Value' AND $\mathbf{P_A}$>= 0.65 AND $\mathbf{C_S}$ = 'Negative') THEN Recommend: "Immediate Salary/Title Review (Urgent)".

Rule 2: IF ($\mathbf{C_V}$ = 'Medium Value' AND $\mathbf{P_A}$>= 0.50 AND $\mathbf{C_S}$ = 'Neutral') THEN Recommend: "Training/Upskilling Opportunity and Manager Check-in (Medium)".

Rule 3: ELSE IF ($\mathbf{P_A}$< 0.3) THEN Recommend: "Standard Retention Monitoring (Low)".

Step 2: Output Generation

2.1. Generate a final action plan: Employee ID, Priority Index $\mathbf{I_P}$, and Recommended Action.

2.2. Collect all structured and unstructured employee data.

Step 3: For each employee:

3.1. Execute Algorithm A (Attrition Risk Scorer) to get the Attrition Probability ($\mathbf{P_A}$).

3.2. Execute Algorithm B (Value-Risk Action Generator) using $\mathbf{P_A}$ and computed Employee Value ($\mathbf{C_V}$).

3.3. Execute Algorithm C (Sentiment Scoring) to get the Sentiment Category ($\mathbf{C_S}$).

3.4. Store the resulting Employee ID, $\mathbf{I_P}$ and Recommended Action.

Step 4: Sort the list of results in descending order based on the Priority Index ($\mathbf{I_P}$).

Step 5: Output the Prioritized Action Plan to HR/Management.

Step 6: Stop

4 Case Study

In this section, we have described the steps of the proposed model with the help of a real-life IBM HR Analytics Employee Attrition & Performance dataset collected from Kaggle [20].

4.1. Data Set and Preprocessing

The research utilizes a publicly available, Human Resources dataset from IBM [20], comprising 1,470 employee records and 35 organizational and personal attributes.

Key Features include:

- *Target Variable*: Attrition (Binary: Yes/No).
- *Demographic*: Age, Gender, Marital Status.
- *Organizational*: Department, Job Role, Job Level, Over Time.
- *Satisfaction*: Job Satisfaction, Environment Satisfaction, Relationship with manager Satisfaction, Work Life Balance.

- *Compensation/Tenure*: Monthly Income, Percent Salary Hike, Total Working Years, Years at Company.

Preprocessing involved standard steps:

1. *Feature Encoding*: Categorical features (e.g., Department, Job Role) were converted into a numerical format using one-hot encoding and label encoding.
2. *Feature Scaling*: Numerical features were standardized or normalized (e.g., using Standard Scaler) to ensure no single feature dominates the model training due to its scale.
3. *Data Splitting*: The dataset was split into training and testing sets (80% and 20% respectively) to facilitate cross-validation and evaluation on unseen data.

Given the inherent class imbalance (approximate 16% Attrition rate), techniques like SMOTE or class weighting were applied during model training to minimize bias toward the majority class.

4.2. Phase I: Predictive Attrition Modelling (Quantitative)

The goal of this phase was to identify the optimal model for predicting attrition risk. A set of classification algorithms was tested using **5-fold cross-validation** on the training data.

Models Evaluated:

- Logistic Regression (LR): Basic model for understanding.
- K-Nearest Neighbors (KNN): Distance-based non-linear classifier.
- Decision Tree Classifier (DTC): Highly understandable model for feature importance extraction.
- Support Vector Machine (SVM): It is useful for binary classification.
- Random Forest Classifier (RFC): An overall method reducing overfitting and improving accuracy.
- XGBoost: A highly optimized gradient boosting framework known for industry-leading performance on structured data.
- Naive Bayes: Models the distribution of inputs of a given class or category.

Logistic Regression

Cross-Validation Scores: [0.86864407 0.88085106 0.89787234 0.89361702 0.89361702]

Mean Accuracy: 0.886920302921024

K Neighbours Classifier

Cross-Validation Scores: [0.83474576 0.84680851 0.86382979 0.84255319 0.85531915]

Mean Accuracy: 0.8486512802019472

SVM

Cross-Validation Scores: [0.86016949 0.86808511 0.86382979 0.87234043 0.87234043]

Mean Accuracy: 0.867353047241255

Random Forest Classifier

Cross-Validation Scores: [0.86440678 0.87659574 0.86808511 0.86382979 0.83829787]

Mean Accuracy: 0.862243058059863

Decision Tree Classifier

Cross-Validation Scores: [0.75423729 0.77446809 0.83404255 0.77021277 0.8]

Mean Accuracy: 0.7865921384781824

XG Boost

Cross-Validation Scores: [0.84745763 0.87659574 0.87659574 0.86382979 0.85957447]

Mean Accuracy: 0.8648106743598991

Model selection was primarily based on the cross-validated Mean Accuracy, with a focus on metrics relevant to imbalanced data, such as F1-Score and Area Under the Curve (AUC). The **XGBoost** and **Random Forest** models consistently demonstrated the highest performance, indicating the non-linear nature of attrition drivers.

4.3. Phase II: Employee Value Scoring and Compensation Simulation

This phase shifts the focus from prediction to prescription by defining a quantifiable measure of an employee's criticality, the Value of Employee(V_E) .

The proposed metric combines objective performance and organizational factors. Employees were then segmented into Importance Levels (e.g., High, Medium, Low) based on their (V_E) score.

This value score was then combined with the attrition risk (R_A) derived from the best-performing ML model to inform a Targeted Compensation Strategy. For employees classified as High Value and High Attrition Risk, a specific Recommended Hike Percentage was simulated as shown in equation-1.

$H\% = f(V_E, R_A, M)$ ………. (1)

Where M is the current Monthly Income, ensuring the recommended compensation is meaningful in the context of current earnings and the risk/value profile. This process provides a prescriptive output, translating a model prediction into an actionable, budget-conscious intervention. The hike percentage was typically simulated to be between 15% to 25% for the highest-risk/highest-value segment.

4.4. Phase III: Sentiment Analysis Integration

To understand the contextual **'Why'** behind the attrition risk, an NLP layer [18] was integrated using a modern machine learning model as follows.

1. **Model Selection:** A pre-trained BERT (Bidirectional Encoder Representations from Transformers) model, fine-tuned on general sentiment analysis tasks, was adapted to analyze hypothetical open-ended responses to typical employee survey questions (e.g., satisfaction with job role, manager support, career growth, work environment).

2. **Input Simulation:** Since the quantitative HR dataset does not contain textual responses, a model was developed to ingest multi-part textual feedback (representing an employee survey) and output a normalized sentiment score.

3. **Sentiment Scoring:** The model processes each answer and outputs a score (ranging from -1 for highly negative to +1 for highly positive) and an associated confidence score. An Overall Sentiment (Negative, Neutral, Positive) is then derived by averaging the individual scores.

4. **Integration:** The generated Overall_Sentiment metric is then concatenated with the dataset generated in Phase II. The hypothesis is that a Negative Sentiment combined with a High Attrition Risk indicates a voluntary exit motivated by cultural or managerial dissatisfaction, warranting non-compensation-based interventions (e.g., management training, better work-life balance policies). A Positive Sentiment with High Attrition Risk suggests an external factor (e.g., better pay elsewhere), which is where the Phase II compensation simulation becomes most relevant.

Therefore, the confluence of the three analytical phases creates a robust, prescriptive HR framework, moving from **"Who is leaving?"** to **"Who is valuable, why are they leaving, and what should we do about it?"**

The final employee risk profile integrates machine learning predictions with sentiment analysis to produce an adaptive retention indicator.

The framework creates a Specific Intervention Matrix, which serves as the actionable foundation of the system. The operational pipeline includes continuous processing of structured HR data and unstructured text (feedback, surveys), regular model scoring through XGBoost and BERT, and automated dashboards that classify employees by engagement and risk levels. The high-risk employees are flagged for immediate HR action. Interventions are given priority. Such as workload and engagement programs for sentiment-related issues. Salary adjustments for market-driven risks. This guarantee focused, economical retention tactics and optimizes organizational return on investment.

5 Results and Discussion

In this section, we have evaluated results obtained from the case study and explained their applicability for employee attrition riskmanagement. The full code is publicly available in our GitHub link [21]: https://github.com/sujaldas8777/Employee_Attrition_and_Salary_Forecasting.

5.1. Attrition Model Performance

The evaluation of the ML classifiers revealed significant performance differences, with ensemble and boosting methods outperforming the linear and instance-based models are shown in figures 2, 3, 4 and 5 respectively.

Classification Report (Random Forest):

	precision	recall	f1-score	support
0	0.87	0.96	0.92	247
1	0.57	0.28	0.37	47
accuracy			0.85	294
macro avg	0.72	0.62	0.64	294
weighted avg	0.83	0.85	0.83	294

Fig.2: Classification report for Random Forest Model

First 10 Predictions vs Actual (Random Forest):

	Actual	Predicted
0	0	1
1	0	0
2	0	0
3	0	0
4	1	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	1

Testing Accuracy (Random Forest): 0.8503

Fig.3: Testing accuracy of Random Forest Model

Classification Report:

	precision	recall	f1-score	support
0	0.88	0.96	0.92	247
1	0.59	0.34	0.43	47
accuracy			0.86	294
macro avg	0.74	0.65	0.68	294
weighted avg	0.84	0.86	0.84	294

Fig.4: Classification report for XG Boost Model

First 10 Predictions vs Actual:

	Actual	Predicted
0	0	1
1	0	0
2	0	0
3	0	0
4	1	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	1

Testing Accuracy: 0.8571

Fig.5: Testing accuracy of XG Boost Model

The superior performance of XG Boost and Random Forest underscores that the drivers of attrition are highly non-linear and involve complex feature interactions. The trial of Random Forest and XG Boost showed testing accuracy nearly 86%, which likely represents the true, generalized performance achievable with proper cross-validation and hyper-parameter tuning, which is an acceptable result for a real-world HR problem.

5.2. Feature Importance: Drivers of Attrition

The feature importance analysis is extracted from the models like Random Forest and XGBoost, provided crucial insights into organizational options such as:

1. **Over Time:** It consistently ranked as the most significant predictor of attrition. Employees reporting working overtime had a relatively high likelihood of leaving. This suggests a direct correlation between work-life imbalance and turnover.
2. **Job Level / Monthly Income:** These characters were also significant. And they have a strong correlation. A major turnover risk was associated with lower job levels and monthly incomes. This highlights the critical role that pay and career path play in retention.
3. **Job Satisfaction and Environment Satisfaction:** These features were also crucial. Problem with the immediate work environment or the specific role is solvable.
4. **Years At Company:** A critical non-linear relationship was observed. Attrition risk is highest during the 1-3 year mark (freshman) and again around the 8-10 year mark (career peak).

5.3. Employee Value and Salary Hike Recommendations

The Phase II analysis using the V_E score yielded a strategic segmentation of the workforce [Fig 6].

	EmployeeNumber	Value_of_Employee
0	1	5.91
1	2	5.61
2	3	7.35
3	4	7.44
4	5	4.98
...	...	...
1465	1466	5.84
1466	1467	5.31
1467	1468	5.91
1468	1469	5.86
1469	1470	4.87

Fig 6: Value_of_Employee (1 to 10 scale) for all Employees

	Value_Emp	Attrition	Income	Hike_Pct	New_Income
1352	5.36	0	5033	0.00	5033.0000
118	4.26	0	2835	0.00	2835.0000
41	5.98	0	2341	0.00	2341.0000
773	6.86	0	8858	0.00	8858.0000
1236	6.01	1	6134	7.76	6609.9984
836	4.63	1	7336	0.00	7336.0000
132	7.00	1	4559	13.87	5191.3333
767	4.66	0	4107	0.00	4107.0000
992	4.54	0	10920	0.00	10920.0000
1279	6.09	1	2342	12.12	2625.8504

Fig 7: Salary hike according to Employee Value and Attrition

Compensation Simulation: The analysis helps to identify a critical group of high-value employees who show high attrition risk. For this segment, the simulation recommended an average hike percentage of approx. 15%(with a standard deviation indicating a range of 10% to 20%), which is notably higher than the typical annual salary increase [Fig 7]. This targeted approach ensures that budget is allocated strategically to secure the most critical talent. The output of this phase (e.g., Recommended_Hike and New_MonthlyIncome) directly enables budget planning for retention initiative.

5.4. Sentiment Analysis Insights

The sentiment integration provided the qualitative context necessary for effective intervention planning. The overall sentiment distribution across the simulated employee responses showed a majority categorized as Positive or Neutral. However, the key insight emerged from cross-referencing sentiment with attrition risk:

- **Negative Sentiment + High Risk:** These employees are likely experiencing immediate stressors or dissatisfaction (e.g., poor communication, high workload /breaks). The primary intervention here should be non-monetary—a management review, workload rebalancing, or psychological support—as a salary increase may only postpone, rather than resolve, the underlying cause.

- **Positive/Neutral Sentiment + High Risk:** These employees are likely leaving for external reasons (e.g., a better offer, relocation, or a perceived lack of internal opportunity despite current satisfaction). This is the prime target for the Targeted Salary Hike and career path clarification.

```
SUMMARY STATISTICS
=====================================================

Overall Sentiment Distribution:
   NEUTRAL      :  52 employees ( 55.9%)
   POSITIVE     :  30 employees ( 32.3%)
   NEGATIVE     :  11 employees ( 11.8%)

Average Sentiment Score: 0.1215 (range: -1 to +1)
Average Confidence Score: 0.6388 (range: 0 to 1)
```

Fig 8: Sentiment Scores

Preview of Results:

	Employee_id	Overall_Sentiment	Avg_Sentiment_Score	Avg_Confidence_Score
0	3	POSITIVE	0.4	0.7543
1	15	POSITIVE	0.3	0.6320
2	34	NEUTRAL	0.1	0.6297
3	46	POSITIVE	0.3	0.6717
4	51	NEUTRAL	-0.2	0.6473
5	52	POSITIVE	0.6	0.6816
6	90	POSITIVE	0.5	0.6865
7	101	POSITIVE	0.5	0.6425
8	103	NEGATIVE	-0.4	0.6471
9	108	NEUTRAL	0.2	0.6793

Fig 9: Sentiment results for first 10 employees

The average sentiment score of all analysed employees was approximately 0.40, which is on a scale from -1 to +1, that suggest an overall positive workplace environment. However, the BERT model provides more precise and actionable insights than traditional binary attrition models, as it can detect localized negative sentiments within specific survey responses—for instance, identifying an employee who is satisfied with their role but holds an unfavourable view of their manager, as shown in Figure 8 and Figure 9.

6 Comparative Analysis

In this section, we compare our proposed work with several established models to highlight the novelty and effectiveness of our model. In the evaluation table-1, **YES** indicates that the parameter is explicitly addressed in the document. **NO** signifies that the parameter is not addressed at all. **PARTIAL** denotes that the parameter is indirectly addressed or inferred through a limited modality. For example, when the model uses HR data but not Sentimental analysis.

MODEL	METHOD USED	INCLUDES SENTIMENT ANALYSIS	CAPTURE NON-LINEAR RELATIONSHIP	EMPLOYEE VALUE RATING (CRITIC)	INTEGRATES STRUCTURED + UNSTRUCTURED DATA	PROVIDES ACTIONABLE INTERVENTIONS
Liu, Y., & Zhang, J., 2021 [7]	Random Forest	NO	YES	NO	YES	NO
Mohiuddin et al, 2023 [1]	XG Boost + GPT-4 (for text output)	NO	YES	NO	NO	NO
Pawar et al., 2024 [9]	XG Boost + VADER/BERT	YES	YES	NO	PARTIAL	NO
Mouli et al, 2023 [10]	Bi-GRU + BERT	YES	NO	NO	NO	NO
Manafi Varkiani, et al. , 2025 [11]	Random Forest + SHAP	NO	YES	NO	NO	NO
UARS (Proposed)	**XG Boost+ BERT + CRITIC**	**YES**	**YES**	**YES**	**YES**	**YES**

Table 1. Comparative Analysis with Other Established Models

8 Conclusion and Future Work

This research work effectively combines quantitative predictive models with sentiment analysis based on deep learning to create a novel framework for employee retention. By analysing the real-world dataset, this research established that non-linear models work well for predicting attrition and structural factors like over-time, job level and monthly income are crucial drivers. Most importantly, this research introduces a measurement for employee value metric that helps organizations for their employee retention efforts where they will have the most impact, supporting significant compensation adjustments. The BERT-based sentiment layer also converts open ended text data into a diagnostic tool, that allows HR to determine the appropriate type of intervention, whether financial or environmental, to tackle the root cause of employee attrition. Hence, the proposed three-phase model provides a practical framework for changing HR analytics from basic risk reporting to a more integrated strategy engine focused on value.

In future, this work can be extended on areas like multimodal integration such as analysing audio from meeting transcripts or studying the flow of internal communications, to deepen our understanding of the workplace. Creating a reinforcement learning system to adjust the weighting factor in the final score dynamically could be another extension of this research work.

References

1. Mohiuddin K., Alam M.A., Alam M.M., Welke P., Martin M., Lehmann J., & Vahdati S. (2023). Retention Is All You Need: An HR-Decision Support System (HR-DSS) using Explainable AI for Employee Attrition. In Proceedings of CIKM ’23. DOI:10.1145/3583780.3615497.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805
3. Glebe, J., Kühn, S., & Schmitz, C. (2019). Strategic compensation management and employee retention: A predictive analytics approach. *Human Resource Management Review, 29*(4), 100691. https://doi.org/10.1016/j.hrmr.2018.11.001
4. Jain, A., Nair, S., & Nair, R. (2020). Predicting employee attrition using machine learning algorithms. *International Journal of Advanced Computer Science and Applications, 11*(8), 104–111. https://doi.org/10.14569/IJACSA.2020.0110813
5. Kim, S., & Kang, H. (2021). Sentiment analysis in human resources: A review and future research directions. *Applied Sciences, 11*(5), 2103. https://doi.org/10.3390/app11052103
6. Lee, J., & Huang, Y. (2020). Using deep learning to predict employee turnover: Integrating textual feedback and numerical HR data. *Computers in Human Behavior, 112*, 106472. https://doi.org/10.1016/j.chb.2020.106472
7. Liu, Y., & Zhang, J. (2021). Employee attrition prediction using explainable AI and ensemble methods. *Expert Systems with Applications, 167*, 114153. https://doi.org/10.1016/j.eswa.2020.114153
8. P. Ghosh, S. Das, S. Roy, A. Bhattacharjee, A. Cortesi and S. Sen, (2025) "SentiTSMixer: A Specific Model for Sales Forecasting Using Sentiment Analysis of Customer”, *IEEE Access*, vol. 13, pp. 85882-85897, doi: 10.1109/ACCESS.2025.3570080.
9. Pawar, Roshini V., Ijantkar, Varad R., Bhandari, Shruti S., Madansure, Dinesh M., & Jadhav, Sakshi S. (2024). “Proposal for employee attrition prediction system and sentiment analysis with GUI development”, *International Research Journal of Modernization in Engineering, Technology and Science* (IRJMETS), 6(11), 556–561.
10. Mouli, N., Das, P., Bin Muquith, M., Biswas, A., & Kabir Niloy, M. D. (2023). Sentiment analysis to determine employee job satisfaction using machine learning

techniques [Undergraduate thesis, BRAC University]. BRAC University Institutional Repository. https://dspace.bracu.ac.bd/xmlui/handle/10361/19357.

11. Shahin Manafi Varkiani, Francesco Pattarin, Tommaso Fabbri, Gualtiero Fantoni (2025), Predicting employee attrition and explaining its determinants, *Expert Systems with Applications*, https://doi.org/10.1016/j.eswa.2025.126575.
12. Ma, L., & Ye, X. (2019). Improving employee satisfaction analysis through BERT-based sentiment models. *IEEE Access, 7*, 145034–145045. https://doi.org/10.1109/ACCESS.2019.2945061
13. P. Ghosh, O. Samanta, T. Goto and S. Sen, (2024) "Sales Forecasting of Overrated Products: Fine Tuning of Customer's Rating by Integrating Sentiment Analysis", *IEEE Access*, vol. 12, pp. 69578-69592, doi: 10.1109/ACCESS.2024.3402133.
14. Wang, L., Zhang, D., & Zhao, H. (2020). A hybrid XGBoost framework for employee churn prediction. *Procedia Computer Science, 174*, 244–253. https://doi.org/10.1016/j.procs.2020.06.045.
15. Subhankar Roy, Arnab Charit, Mriganka Patra, Ananya Sadhukhan, Diya Chakraborty, Partha Ghosh, Kingshuk Chatterjee, Anirban Mukhopadhyay, (2025), "A study of the cutting-edge general-purpose compressors' performance on the normalized genome sequence", *Gene Reports*, Volume 41, https://doi.org/10.1016/j.genrep.2025.102358.
16. Sun, S., Li, C., & Xu, J. (2018). Predicting employee turnover with machine learning algorithms: A case study in China. *International Journal of Forecasting, 34*(4), 1218–1230. https://doi.org/10.1016/j.ijforecast.2018.03.011
17. Tzeng, G.-H., Chiang, C.-H., & Li, C.-W. (2017). Predicting employee turnover using decision trees and neural networks. *Human Systems Management, 36*(2), 115–126. https://doi.org/10.3233/HSM-17155
18. Ghosh, P., Charit, A., Banerjee, H. et al. (2025), "DropWrap: A Neural Network Based Automated Model for Managing Student Dropout", *International Journal of Networked and Distributed Computing*, 13, 17. https://doi.org/10.1007/s44227-025-00058-z
19. P. Ghosh, S. Som, S. Sen. (2018) "Business Intelligence Development by Analysing Customer Sentiment", International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), doi: 10.1109/ICRITO.2018.8748517.
20. Data Link: "IBM HR Analytics Employee Attrition & Performance", https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset, Accessed on 25/10/2025.
21. GitHub link for code: https://github.com/sujaldas8777/Employee_Attrition_and_Salary_Forecasting.Accessed on 25/10/2025.
22. Yadav, S., & Pal, S. (2022). "Application of explainable machine learning in human resource analytics: Employee attrition prediction". *Information Systems Frontiers, 24*(3), 899–915. doi: 10.1007/s10796-021-10139-8.

.