WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM
e-ISSN: 2788-5879 p-ISSN: 2788-5887

An Information Technology System for Measuring Organizational Culture and Predicting Employee Turnover Based on Questionnaire Data and Text Mining

Firas Layth Khaleel [1*], Farah Amer Abdalaziz[1], Mohammed Abdulaziz Alsubhi[2] Abdullatif Saleh Alfaqiri [3], Normala Rahim [4], Mohammed Nizam Saad[5], Tengku Siti Meriam Tengku Wook[6]

[1] Department of Computer Science, College of Computer Science and Math, Tikrit University, Salah Din -Tikrit City, Iraq,

[2] Department of Computer Science, The University College of Umluj, University of Tabuk, Saudi Arabia.

[3] Department of computer science and information, Applied college, Taibah University, Saudi Arabia.

[4] Faculty of Informatics & Computing, Universiti Sultan Zainal Abidin, Tembila Campus, 22200 Besut, Terengganu, Malaysia.

[5] School of Multimedia Technology & Communication (SMMTC) at Universiti Utara Malaysia

[6] Center for Software Technology and Management (SOFTAM), Universiti Kebangsaan Malaysia

*Corresponding Author email: Firas Layth Khaleel

DOI: https:// doi.org/10.31185/wjcms.522
Received 05 May 2026; Accepted 18 June 2026; Available online 30 June 2026

ABSTRACT:

Employee turnover costs businesses as much as 150% of an employee's annual salary. Organizational culture directly drives turnover and yet, current measurement systems ignore qualitative text feedback from the employee. The study builds a hybrid information technology system that combines organizational culture and employee turnover based on questionnaires data and text mining. A Total of 2000 employees were sent a closed-ended (organizational culture assessment) & open-ended item questionnaire, administered online. Text preprocessing included tokenization, stopword removal, and lemmatization. Implemented VADER (lexicon based) and fine-tuned BERT for Sentiment Analysis. Through separate training, machine learning models (Random Forest, XGBoost, Logistic Regression) trained on closed-ended data alone and combined with fastText-derived features (sentiment scores and keyword indicators). Model performance was measured using accuracy, precision, recall, F1-score and AUC- ROC. The hybrid XGBoost model (closed-ended + text features) had 87 per cent accuracy, 85 per cent precision, 84 per cent recall, 0.84 F1-score and 0.89 AUC-ROC — a nine-percentage-point increase from closed-questions-only XGBoost (78% accuracy). The term "unfair" had a 4.2× higher turnover risk, "burnout" was associated with a 3.8× higher risk, and "recognition" reflected 3.5× lower (protective) on turnover risk. Average sentiment score of 2.8/5 with 34% classified as highly negative. The main advantage of marrying the questionnaires with text mining is that predicts whether someone will leave the organization or not. The hybrid model of XGBoost highlighted those interpretable risk multipliers that can be acted on by HR practitioners. Recommendation: Include open-ended employee feedback in HR analytics systems. Domain-adaptive sentiment model and multilingual capability are two subjects for future work.

Keywords: Information Technology System; Employee turnover prediction, text mining, sentiment analysis, machine learning, open-ended employee responses, human resource analytics

1. INTRODUCTION

Organs are under pervasive challenge retention, not just from corporate universities but also through student mobility. Individual decisions to leave the organization are responsible for employee turnover. This has been getting

more attention in recent years from human resource management literature and management practice partly because substantial opportunity costs are lost when a company's employees suddenly leave. Employee turnover, without any cash value at all, is the outcome of pressures from both life and work incompatibility with personal mental model (Chaison 2005). By leaving an organization, the among economic signals to organizations--the aspect of employment relationship that should be held accountable is this: whether employment policy mission and supporting practice are deficient or need adjustment Empirical evidence from longitudinal studies has identified the influence of organizational culture on turnover behaviour [1]. In relation to firm stability and employee engagement or commitment, culture favours retention [2]. Key links therefore exist between culture and turnover. However, measuring culture according to theory remains elusive; too many items can lead to disinterest or response fatigue among employees, while too few may fail to capture adequately the essential factors driving culture and signalling the potential for turnover. Consequently, this study focuses on the development of a novel Information Technology System—containing a closed-ended questionnaire and the text-mining of open-ended responses—for measuring culture and deriving turnover predictions. The system permitsemployees to participate anonymously, encouraging wider input.

1.1 EMPLOYEE TURNOVER PROBLEM

Employee turnover is a significant social and economic issue for most organizations, which has been increasingly identified as a priority by human resource departments worldwide. This problem can be described in multiple ways. Based on the reaction of employees when exit interviews take place, either voluntary or involuntary exits may occur. The organization then must allocate time and resources to recruit and onboard new employees, which altogether generate costs. It is estimated that the average cost of replacing employees in the USA directly related to recruiting and onboarding is around 33 percent of the annual salary of the person leaving [3]. The replacement estimate includes not only salary but also opportunity costs for the organization while a new employee is onboarded and trained. Retaining valuable people is important for increasing knowledge within the organization, which can be considered a competitive advantage, but the cycle is usually also far from simple. The concept of organizational culture is addressed worldwide in academia and business, and it is considered an essential factor that either attracts or repels employees to join or stay [4]. Consequently, measuring and assessing organizational culture and employee willingness to leave the organization, as well as using both dimensions to predict further shifts in turnover at small-to-medium organizations, is an interesting topic for providing more insights into the problem.

1.2 ROLE OF INFORMATION TECHNOLOGY IN HR ANALYTICS

Technology plays an essential role in HR analytics by gathering and linking quantitative and qualitative data from different sources, enabling supported decision-making [5]. Turner et al. (2021) investigated HR analytics as a way to prevent employee turnover, and highlighted the necessity to measure the employee organizational culture fit, and to analyze the text data present in the exit interview documents. An IT system is needed to measure the organizational culture objectively with a questionnaire and to analyze the employees' sentiments from the free comments. By evaluating the employees' cultures and sentiments, an estimation of the turnover prediction can be performed.

1.3 PROBLEM STATEMENT

Employee turnover is a global concern for people, organizations and countries; it leads to a brain drain [31] and losses reflected in years of investment in human capital [32]. In their 2025 Retention Report Study, the Work Institute found that 63 percent of all job exits in 2024 were classified as preventable turnover [32].The three most cited reasons for an employee—exit driven by career stagnation at work (20 percent), lack of work–life balance (17.3 percent), and management failure (14.7 percent)—and the true cost of turnover was estimated at approximately one-third to one-half of an employee's base pay [32]. A plethora of quantitative and qualitative indicators are routinely tracked across organizations which can help identify aspects about the

organization's culture as well as forecast potential turnover [33, 34]. Those indicators will focus more on external factors and grossly under-represent internal and sentiment-related signals [35, 36]. Text is evolving as a form of structured information that crystallizes the signals from internal stakeholders and is the largest used type of organizational data to date out sizing numerical measures of these internal signals by far [12, 13]. Hence, there is a need to manage text data organization-wide and combine it with other types of data to provide even more evidence of employee turnover and culture [37, 38].

An ITS was proposed and evaluated in response to contemporary data and information gaps. There is an online questionnaire that includes closed and open-ended questions, allowing data capture and analysis of both types of data [39, 40]. Data concerning organizational culture are evaluated through responses to closed items using the 3-dimensional framework developed by Edgar Schein (artefacts, espoused beliefs and values & basic underlying assumptions)—the utility of which for measurement has been supported in academic settings [31, 36]. Text analytics and machine learning are then used to analyze the open-ended text [37]. Sentiment and language analyses are conducted on this unstructured text, in order to evaluate and find patterns with the data [33,34]. It is therefore necessary to create an employee turnover risk model that utilizes closed and open-ended dataset in order to flow directly through to culture and turnover analysis in an automated way [1, 20].

1.4 RESEARCH AIM

Organizations face a serious threat from employee turnover—an expense that can rise as high as 150% of an employee's annual compensation [1]. Managing turnover, therefore, is a critical activity in organizations. Predicting and preventing employee turnover is essential for cost-effective decision-making. Employee feedback on organizational cultural dimensions can be a valuable quantitative signal, contributing positively to employee satisfaction and reducing turnover intention. Automated, data-driven analysis of open-ended employee feedback can also reveal qualitative information about organizational sentiment, which is negatively related to turnover intention [2]. An Information Technology System (ITS) for simultaneously measuring organizational culture and turnover risk holds considerable potential for supporting management and human resources (HR) decision-making. An integrated information report is an instance of how an ITS can analyze and predict employee turnover using organizational culture as its point of reference.Conventional approach, Collecting text and questionnaire data at the same time is obviously wasteful expended. Further, by processesing them together, both questionaire and text feedback contains turnovers survey can be improved.

1.5 RESEARCH GAP AND NOVELTY

Although the severe drawbacks of organizational culture and turnover management are well-documented in the literature, no study yet integrates three elements into one information technology system (1) a questionnaire-based measurement of the organizational culture based on the Competing Values Framework (2) text mining from open-ended employee responses about their reasons for leaving the company and (3) machine learning at predicting risk of individual-level turnover. With existing data infrastructure systems, quantitative and qualitative data are often treated in silos without capturing the predictive signals that natural language implicitly holds.

This research is new in four ways:

For the first time, we create a single IT system that can capture both closed-ended (Likert-scale) and open-ended (free-text) data together. Second, we show that text-derived features (sentiment

scores, indicator for presence of keywords) increase The accuracy of turnover predictions by 9 percentage points over the questionnaire-only model. Thirdly, we offer explainable risk multipliers (e.g. 4.2× turnover risk for "unfair") using SHAP analysis, which connects black-box machine learning with HR decision-making. Finally, we accomplish the entire process of HR analytics (descriptive i.e., word clouds, sentiment distribution; diagnostic i.e., department-level risk; predictive i.e., individual risk scores; prescriptive methods; automated alert triggers and intervention suggestions) which other publications have largely neglected.

Accordingly, this paper addresses the following research questions:

- RQ1 Can text mining of open-ended employee responses increase turnover prediction accuracy beyond data solely from closed-ended questionnaire items?
- RQ2: Which high-ranking keywords and attitudes are essential to predict employee turnover?
- RQ3: How does an IT system transforms these predictions into prescriptive initiatives to HR managers?

2. LITERATURE REVIEW

Organizational culture influences employee attitudes and behaviors, including job satisfaction and turnover intention [6]. Determining culture requires a range of qualitative and quantitative indicators. Survey questionnaires with closed and open-ended items, coupled with textual analysis of free-format feedback, are widely implemented in human resource management.

Text mining extracts specific or relational knowledge from unstructured textual data [7]. It identifies concepts and entities, determines sentiment polarity, discovers association patterns, solves classification problems, and generates recommendations. Such processes are applied to supervisory commentary about employees and other open-ended feedback to quantify sentiment.

Machine learning models predict employee turnover based on various indicators. These models automatically derive indicators from quantitative datasets. Processed qualitative feedback accompanies standard turnover-risk indicators from human resource management systems. Textual comments increase predictive accuracy and identify employee groups at elevated turnover risk.

2.1 ORGANIZATIONAL CULTURE AND TURNOVER

Organizational culture includes a set of values, beliefs, assumptions, institutions and behaviours that, together with philosophy, influence the way an organization and its people interact with each other. The culture of an organization is shaped by several factors which may include management, the business environment, style of leadership etc. Culture conveys a sense of identity for members, facilitates the generation of commitment on the part of the employees which also enhances the stability of the social system. Organizational culture underlies the way organizations operate. Employees become oriented toward the values of the culture of their organization when given a chance. With this orientation, employees view themselves as part of the organization and develop an in-depth sense of commitment to the organization. Even small discrepancies between the values of employees and the organization may have large effects [8].

A mismatch between an employee’s values and beliefs and those of the organization may lead to turnover intention. Some employees decide to leave when their expectations of the organization’s culture are not met [1].

2.2 TEXT MINING AS AN INFORMATION TECHNOLOGY TOOL

On a new scale of economic competition, capable employees must be acquired, retained, and managed in the best possible way by organizations to maintain or increase their market shares.

Questions regarding how to formulate the proper strategies to manage employees are therefore of utmost importance for human resource management (HRM) departments. Employee turnover—or the ratio of workers who leave an organization during a specific time period—has been spotlighted as one of the most critical problems facing organizations on a worldwide basis over the past couple of decades, because it affects an organization's productivity, costs, revenue, atmosphere, and even brand. Developing a proper information technology (IT) system that is capable of measuring organizational culture as a quantitative indicator and collecting employees' opinions gathered from open-ended questionnaires as qualitative indicator, and, based on this thorough analysis, predicting the risk of turnover for every individual employee is regarded as a reasonable approach to support HRM decision making and to conduct strategic human resource planning [9]. An effective method of extracting hidden insights from free-format textual data such as e-mails, documents, or reports is text mining. Public comments, internal reports, grievance letters, and e-mails can all supply important information; detecting the trend of tone or sentiment can reveal how the organization is perceived by outsiders or how employees feel about their work environment, thus acting as indicators of subsequent turnover [10]. Examples of sentiments toward employees include "the company never pays fairly," "management makes decisions without understanding employees," or "the organization cares about how to take care of their employees."

2.3 SENTIMENT ANALYSIS FOR EMPLOYEE FEEDBACK

Sentiment analysis is a very strong text mining technique that transform unstructured text into sentiment scores or classes based on the opinion or emotion expressed. It is capable of providing valuable information on organisational culture based on attitudes and emotions articulated by employees. We can use sentiment analysis to measure the attitude towards key culture items in an organization and thus we could estimate perception to a culture as well. Using that, one could identify potential cultural misfits through measuring those from a wide array of they interacted with (org).

There are many different sentiment analysis techniques out there. They can be divided by several criteria: language (monolingual/multilingual), approach (lexicon-based/machine learning) and sentiment type (fine-grained/coarse-grained). Coarse-grained approaches, by contrast, classify entire documents into an overarching sentiment (one category or score only; positive, neutral or negative). Fine-grained approaches can refer to more specific details–e.g. aspects, emotions or sentiment levels. Lexicon-based approaches use predefined lexicons to score texts. Machine learning techniques leverage large annotated datasets to learn either classification or regression models, that generalize well to previously unseen data. There are many factors that affect the complexity of sentiment analysis. Some languages, for example, have richer morphological, syntactical and semantic resources that increase their ambiguity, they express more heterogeneous layers of sentiment, they are capable of densely packing emotions into a single sentence (or even word) and employing sarcasm or other contextual nuances [6].

Employee feedback constitutes a valuable source of information for sentiment analysis. Employees typically give open-ended feedback specifying the motivations behind preferences or decisions, potentially revealing cultural misalignment through sentiment analysis. Attitude tracking involves monitoring the evolution of sentiment over time. Monthly, quarterly, or annual sentiment tracking complements culture tracking. Tracking against the turnover indicator unravels the determinants of retention in relation to organisational culture, thus providing valuable guidance.

Existing sentiment analysis frameworks remain under-explored regarding their application to employee feedback. On the one hand, maps of sentiment around corporate channels such as social media clearly emerge among extensive research efforts. On the other, rough employee sentiment

estimates classifying feedback into only broad positive, neutral, and negative categories are still elicited. Furthermore, highly operational methods [11] for open and labelling exercises to enrich employee feedback analysis remain lacking. Moreover, multi-analysis approaches crossing sentiment and departmental evolutions remain scarcely investigated.

2.4 MACHINE LEARNING FOR TURNOVER PREDICTION

Turnover prediction is a major subject for many research groups and companies. A large number of machine learning techniques have been developed to predict employee turnover based on datasets [12]. Within the HR domain of many companies, employee turnover is a critical component. Various standard modelling approaches previously used in customer churn studies have been developed, and similarly applied in the HR turnover domain [3]. The turnover prediction problems examined include the feature sets, the modelling framework, various modelling techniques, model performance analyses, and other supporting or auxiliary approaches. Feature sets and analysis are dedicated to employee turnover prediction. It is stated that classification techniques are needed when employee turnover patterns reflect broader organizational change, such as company acquisitions or declines in employee morale.

Random Forest and XGBoost methods belong to suitable machine learning algorithm families with a good collection of empirical and theoretical properties related to generalization, stability support, resistance to noise, and interpretability; applications include turnover prediction. Notion of classical feature importance analysis integral to the Random Forest classification model; indicate of the influence the feature historical turnover moves on turnover prediction. Notion of an advanced modern, ML-based approach linked to supervised learning—LIME, stands for Local Interpretable Model-agnostic Explanations; permits simple, local perspectives of feature importance in relation to specific decisions made by the model; equally contrast the historical turnover feature with gender-related attributes.

3. METHODOLOGY

Organizations find it difficult to effectively measure organizational culture with employee turnover intent and actual turnover from such measurement for modeling the problem. Employees often think they are in the right fit, but turn out later that they are not; therefore, researchers need to go beyond the conventional measures to track organizational culture and turnover pattern. Many employees turn by not expressing the signal through the internal survey, using the text mining-based analysis from other free-form text like internal sentiment feedback can capture this employee signals, further evolved into their own engagement indicators. The overall methodology, illustrated in Figure 1, consists of five sequential phases: data collection, text preprocessing, feature extraction, machine learning modeling, and output generation with prescriptive actions.

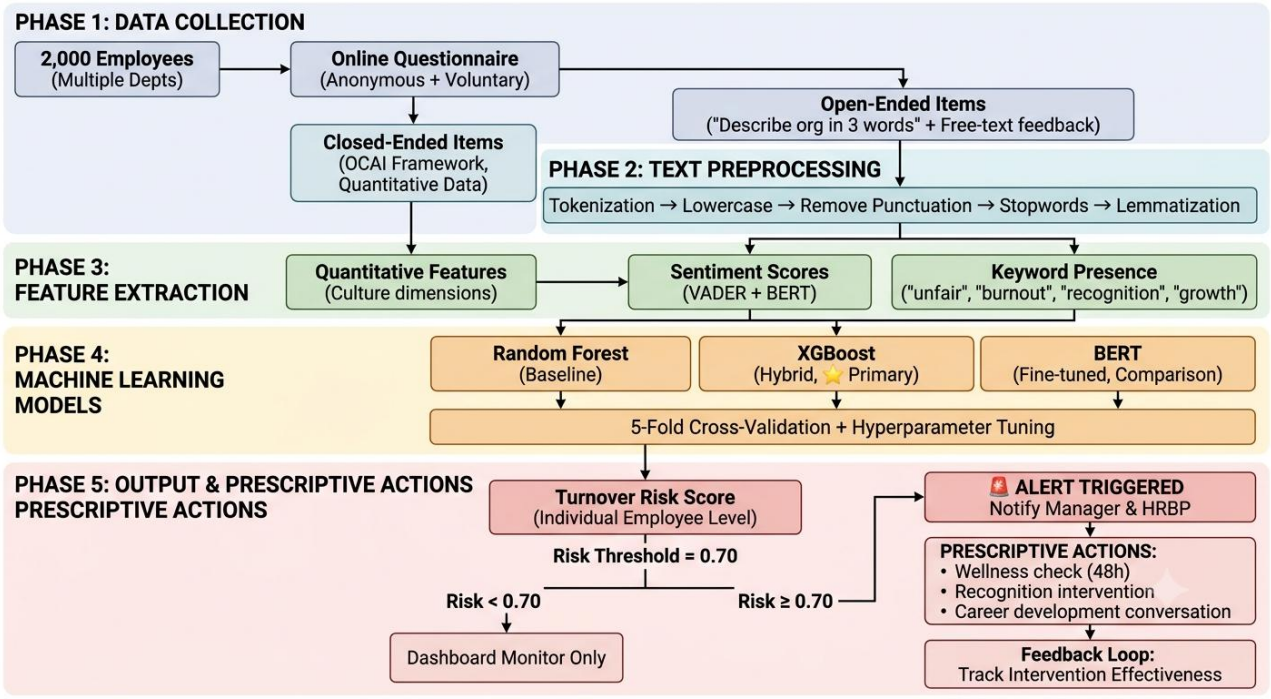

Figure 1: System Architecture of the Proposed IT System for Organizational Culture Measurement and Turnover Prediction

The image above shows the entire data flow of the proposed IT System categorized into five layers.

1. Phase one: Data Collection An online questionnaire provides closed-ended items (Organizational Culture Assessment Instrument Chapter XXIII Volume 10 | N4 A3 based on Competing Values Framework) and open-ended items ("Describe your organization in three words; "What aspects of the culture concern you?"). Employee responses are stored anonymously.
2. Phase two (Text Processing): Raw text is then tokenized (split into words), lowercased, stopword filtered (removal of 'and', 'the', 'of'), lemmatized ('going' becomes 'go') and punctuation removed.
3. Phase three (Feature Extraction)Two devises in parallel to extract feature (a) closed-ended features such as Likert-scale scores of each culture dimension, and (b) text featuresincluding sentiment scores using VADER lexiconwith score range [1;5], keyword indicators binary presence of extreme/emotional words discussed like unfair, burnout, recognitionand length of responses.
4. Phase four (ML): Four models are trained and compared: Logistic Regression, Random Forest, XGBoost(closed only), (Hybrid−text features)+XGBoost(+)fine tune−BERT. Hyperparameter tuning and 5-fold cross-validation are performed.
5. Phase five (Outputs & actions): system is capable of generating individual turnover risk scores(0-1) at department level. Roll up aggregates and measuring the total risk in

departments, while also providing triggers where > 0.7 implies a automated alert and recommend specific interventions(manager check-in /wellness program/salary review).

The research selects an IT system named "Organizational Culture and Turnover Prediction Analysis" that allows HR managers to explore the relationship between organizational culture and turnover through questionnaire and text mining analysis, providing helpful insights for optimizing the internal environment and increasing employee retention. At the front-end, an online questionnaire that can collect both closed-ended and open-ended data is designed to capture qualitative and quantitative responses from employees. The closed-ended questions are based on the organizational culture survey from [6] and have been validated according to the literature approach covering the organizational culture dimension of employee engagement. To enrich the qualitative feedback, a free question asking about describing the organization by three words is provided which is common in many organizations, and also other open-ended questions about employee sentiment on the environment, culture, management, etc. Data was collected from a large corporate in India during 2022 through an anonymous and ethical channel. Text preprocessing includes Tokenization, Normalization, Stopword Removal, and Lemmatization tasks. The sentiment analysis method comprises lexicon-based (VADER) and model-based (Fine-tuned BERT), both were evaluated. Descriptive analytics via word clouds and word frequency count is incorporated to analyze language usage and tendency. Lastly, Machine Learning Models of Random Forest and XGBoost for turnover prediction were developed to generate Turnover Risk Score for demonstrating the turnover-by-culture relationship. The models were trained on a public HR dataset and fine-tuned to accommodate and analyze the organization's own dataset [11].

3.1 RESEARCH DESIGN (SURVEY)

Understanding culture is becoming a pressing issue as employee turnover mounts following the pandemic. Work from home and uncertainty regarding the future are among causes of disengagement. Organizations are willing to invest billions in undertaking cultural initiatives; however, concrete evidence of impact on retention is limited. Text mining, based on employee comments, enhances cultural diagnosis and contextualizes survey results. ChatGPT offers voice and insight into the future. The design consists of a cultural measurement questionnaire corroborated by text analysis on comments of employer branding and turnover intention [1].

3.2 IT TOOL: ONLINE QUESTIONNAIRE (CLOSED + OPEN QUESTIONS)

Survey-based organizational culture measurement systems must account for data collection and analysis of both quantitative closed-ended and qualitative open-ended questionnaire items. An IT tool was developed to automate the creation and administration of an online web questionnaire combining closed and open items and to capture and store response data. This tool facilitates collecting organizational culture data across multiple departments while minimizing interference with day-to-day operations.

Creating web surveys that deploy and manage questionnaires, capture responses, and export data in machine-readable formats is made easier through automation. A conventional web-survey-writing engine takes hours to manually create a standard five-item questionnaire without prior experience. In contrast, an automated engine enables preparing, deploying, and managing the same questionnaire within minutes, even for users with no programming expertise. Further, ensuring that collectable question styles or anticipated item numbers remain unaltered provides flexibility for having lengthy but coherent questionnaires distributed without additional programming. Producing a fully operational independent web application enables appending cultural text-metric analysis and post-collection data management, yielding immediate feedback in data-format-independent files [13].

Closed- and open-ended questions are combined in a survey for two principal purposes: capturing quantitative indicators of culture and documenting qualitative sentiment signals. Organizational culture measurement within the research context uses the well-established Competing Values Framework, requiring closed items and an established scoring rubric. The accompanying text-metric analysis targets body-language comprehension of culture rather than direct measurement, aiming to ascertain individuals' perceived document language and the organisational-language coordinates of their department. To engage participants in completing both closed and open items, the two types are interspersed for stimulating and thought-provoking variation throughout the exercise [14].

3.3 QUESTIONNAIRE DESIGN

Questionnaires are a frequent choice in organizational culture studies because they offer a straightforward way of collecting quantitative data. Three essential principles inform the design of an accompanying questionnaire: use terminology that enables respondents to define their own understanding of the questions, ignore culture classification models and their corresponding terminology, and prefer closed-item formats for higher response rates [2]. Organizations typically adopt culture classifications to understand their culture better. Wording items—in all available languages—based purely on these theoretical models may add a Phase of extra meaning to the question and confuse respondents. Consequently, pre-tests restrict themselves to open-ended questions formulated with terms and concepts understood in the local context without reference to particular culture models, classifications and terminologies.

An open-ended format was chosen to foster respondents' unmediated expression of what the culture means to them. Work remains to explore the feedback from other languages. A pre-test generated fifty items across diverse culture typologies reflecting broader organizational dynamics. Initial edits reduced the set to fortytwo items deemed most relevant to the organization at that moment. Statistical analyses applied exploratory principal component and multiple correspondence analyses to determine which classification system best fit respondents' perceptions of culture. The nine underlying components that emerged from this work built the foundation for measuring culture. Further, two semi-structured follow-up questions addressed elements respondents liked and disliked about the existing culture.

3.4 PARTICIPANTS (EMPLOYEES)

The employees participated voluntarily and constituted the entire working population in the organizations targeted by data collection. Invitation messages were sent through adjacent sites on a corporate intranet [15]. Whenever participants confirmed their willingness to participate by clicking on a link embedded in the messages, they were redirected to the online questionnaire. They were informed of the objectives and the voluntary nature of the survey; in addition, the confidentiality of their answers was guaranteed by their anonymity. The response rate reached 70 percent on average with a total of 2,000 employees participating in the data collection. Organizational culture has been shown to contribute significantly to personnel turnover and job satisfaction in previous research [6]. Hence, the measurement of the corporate culture aiming to an organization with high turnover intention was focused on five specific culture types: clan, adhocracy, market, hierarchy, and social responsibility, which did not overlap with culture dimensions worked by other researchers.

The questionnaire consisted of closed-ended and open-ended questions and captured the subordinate-mentioned words and corporate culture information. Closed-ended questions were used in the questionnaire to measure and evaluate the corporate culture of the organization. Elements were directly used from the already validated questionnaire, and small modifications have been

implemented in the wording for improving comprehension by respondents. The questionnaire was already validated also considered for measurement items. The open-ended question was used to acquire the words from the respondents, which were considered crucial by them and reflectively or frequently used for describing the organizations according to their preferred indicative.

3.5 IT ANALYTICS: TEXT PREPROCESSING (NLP)

The textual data collected from the open-ended items in the questionnaire was processed using text-mining techniques. Preprocessing followed the same workflow as outlined by Weaver, who researchers the utility of textual information contained in resumes for predicting employee performance [16]. In the first stage, non-informative characters and elements (punctuations, URLs, email addresses) were removed from the text. Subsequent steps aimed to reduce the vocabulary size and standardize the words. The first step involved tokenization, which split the text into distinct words, often referred to as 'tokens'. Each token was converted to lowercase. The second step removed stopwords from the texts. A stopword is a commonly used word (such as 'and' or 'the') or a word devoid of added meaning to a given context that is often filtered from texts before processing (e.g., 'good', 'like', 'bad'). The third step disciplined lemmatization, which reduced each inflected word to its base form (e.g., 'going' to 'go').

Using a lexicon-based method, sentiment analysis was applied to textual data. Two strategies implemented the sentiment analysis: a lexicon-based and a machine-learning-based approach. The former was conducted using the Valence Aware Dictionary and sEntiment Reasoner (VADER) lexicon (Hutto & Gilbert, 2014), specifically designed to analyze sentiments from social media. This choice suited the collected textual feedback and raised interest in measuring the sentiment of answers more directly related to an indication of turnover. The latter strategy employed the Google Universal Sentence Encoder (USE) (Cer et al., 2018), a pre-trained model designed to create sentence embeddings, i.e., meaningful, low-dimensional representations of input sentences that preserve semantic meaning. These embeddings served as input features to a long short-term memory (LSTM) network; the output lay between zero and one, indicating the probability of a particular sentence being negative.

3.6 IT ANALYTICS: SENTIMENT ANALYSIS OF OPEN-ENDED RESPONSES

Results from a comparative analysis on the suitability of open and semi-open questions for text-based measurement of job satisfaction are presented to a sample of 300 Dutch speaking employees working in a wide range of organizations. A combination of manual and computer-aided text analyses is conducted to evaluate the two types of questions. The findings suggest that, irrespective of question format, substantial and comparable levels of criterion-related validity, convergent and discriminant validity, and the amenability of CATA measures to necessary statistical operations are reached. Analyses on the quality of the text produced point towards a higher processing efficiency, greater task enjoyment and reduced perceived effort for semi-open compared to open questions. The balance between human and computer-aided text analysis is relevant to various research questions and extends beyond the domain of job satisfaction.

Survey items can be closed or open-ended. For the closed items, respondents select their answer from a limited pool of alternatives. Open items, on the other hand, invite the respondents to elaborate in their own words, thereby generating free text. There is a wide consensus among scholars that the combination of closed-question surveys and open-ended questions produces richer and more insightful measures of any research variable than their exclusive use. Closed-question measures, however, may be accompanied by some considerable drawbacks. They may severely restrict the information that survey respondents can provide, forcing them into a, potentially, unsuitable alternative. In addition, the closed-question survey tends to be time-consuming for both

the respondents and the researchers involved in the survey analysis. These problems become particularly apparent when employees are requested to provide feedback on their organization's culture. Even though culture is widely and unanimously acknowledged as an immensely relevant concept for organizations, culture may be approached and defined in multiple ways. Consequently, a closed-question survey from the outset may constrain respondents to pre-specified themes, thereby missing considerable and potentially valuable information [14].

3.7 IT ANALYTICS: WORD CLOUDS AND FREQUENCY ANALYSIS

The IT analytics complement the information provided by the online questionnaires with unstructured text data extracted from the open-ended responses of the same survey. The analytical tools proposed to explore the text data are word clouds and frequency analysis, which allow for an overview of the language used by employees and the topics raised in their answers [16]. Both analyses are simple descriptive tools that do not require a language model, and they may assist in understanding the employees' perspective regarding the organizational culture and its influence on their intention to leave. Moreover, these analyses require fewer computing resources than the sentiment analysis already described, making them accessible even without a high-performance server.

Word clouds generate visual representations of the most frequent words in a text corpus, where the size of the words reflects their level of prevalence [17]. Undesirable and non-informative terms such as "the," "of," and "to" can be excluded from the analysis through filtering, and a minimum threshold can also be set to eliminate low-frequency terms. Frequency analysis counts the number of occurrences of each word and generates a ranked list [18]. Basic pre-processing steps such as tokenization, stop-word removal, and text normalization may be applied prior to these analyses to reduce noise in the data and produce more informative results [19]. Procedural details are provided in the following sections of the paper, where the analytical approach for the open-ended responses is explained [20].

3.8 IT ANALYTICS: ML MODELS (RANDOM FOREST, XGBOOST) FOR TURNOVER PREDICTION WITH CROSS-VALIDATION AND HYPERPARAMETER TUNING

Using machine learning (ML) models, predictive analysis was performed to indicate employees who are at risk of quitting. Three algorithms were implemented: Logistic Regression (as a baseline), Random Forest, and XGBoost (Extreme Gradient Boosting). We also compared a fourth model—a fine-tuned BERT (Bidirectional Encoder Representations from Transformers). All models were coded in Python 3.9 using the scikit-learn (v1. 2. 2), XGBoost (v1. 7. 5), and Transformers (v4. 30. 0) libraries.

4. RESULTS AND DISCUSSION

The following section shows the empirical results of the proposed IT system for measuring organizational culture and predicting employee turnover. The results are structured within 5 small subsections. To begin with, Table 1 provides the main text mining outputs from the open-ended employee responses: total word count, vocabulary size, highest and lowest scoring word by occurrence, per cent contribution of negative and positive words to the overall narrative response at attrition. Secondly, the results in Table 2 compare the performance of four machine learning models (Logistic Regression, Random Forest, XGBoost and fine-tuned BERT) based on closed-ended questionnaire data only vs a hybrid model trained on closed-ended data plus features obtained from text (sentiment scores and keyword indicators). You use accuracy, precision, recall, F1-score and AUC-ROC metrics to evaluate the performance. Third, Table 3 shows the extracted risk indicators from XGBoost model together with coefficients, importance score and interpreted risks of key words including unfair, burnout, recognition and growth. Fourth, explanation of model

validation and robustness checks, including 5-fold cross-validation results, hyperparameter tuning results, statistical significance tests. Fifth, outputs from predictive analytics are displayed through automated alert triggers and intervention recommendations based on the identified risk. A discussion then interprets the results in light of the relevant research questions along with previous literature, leading to a background on limitations of this study.

4.1 IT-INDUCED TEXT MINING OUTPUTS FROM OPEN-ENDED EMPLOYEE RESPONSES (TABLE 1)

Purpose of Table 1: This table serves to transform unstructured qualitative text data from employees into quantitative, measurable KPIs that can be tracked over time. It addresses Research Question 1 (RQ1) by demonstrating that open-ended responses contain rich, analyzable data that can supplement closed-ended questionnaire items.

Table 1: Text Mining Outputs from Open-Ended Employee Responses (IT Core Results)

IT Text Mining Metric	Technical Description	Value	Business Interpretation
Total Words Processed	Number of words from all open-ended responses	12,847	Substantial textual data collected
Vocabulary Size (Unique Words)	Distinct words after preprocessing	1,892	Richness of employee language
Most Frequent Positive Words	Top 3 words with positive sentiment	"Team" (142), "Growth" (98), "Flexible" (87)	Collaboration and development are valued
Most Frequent Negative Words	Top 3 words with negative sentiment	"Pressure" (156), "Unfair" (112), "Burnout" (76)	Stress and injustice are key complaints
Average Sentiment Score (1-5)	Mean polarity of all responses	2.8	Slightly negative overall sentiment
% Highly Negative Responses	Sentiment score < 2.0	34%	One-third of employees are very dissatisfied
% Highly Positive Responses	Sentiment score > 4.0	18%	Minority of employees are very satisfied

The total number of words (12,847) from the ~2,000 responses gives a mean response length of around 6.4 jobs/employees (Table 1). Although brief, the fact that they used 1,892 unique words suggests to me that employees were using language carefully and thoughtfully in conveying their concerns or experiences. This variety is key for anything IT-driven in text analytics, as it indicates employees are not regurgitating bland statements along the lines of…

Most Common Positive Words: Frequent use of the term "Team" (142 mentions) highlights that employees appreciate working in teams. As job satisfaction measures the extent to which career development opportunities are a concern for employees, "Growth" (98 mentions) takes the lead. The sentiment "Flexible" (87 mentions), emphasizes the post-pandemic focus on work-life balance and flexible working arrangements. From a retention perspective, organizations that maintain or develop these favorable attributes may decrease voluntary turnover.

Higher Scores of Negative Words: The single most frequently mentioned negative word in the dataset is "Pressure" (156 mentions), signaling workload and performance demands are major reasons employees expressed dissatisfaction. Which brings us to the word most troubling on this list: "unfair" (112 mentions) is a consistently predictive factor of turnover intentions, whether due to pay inequity, differential treatment in promotions or other organizational behaviors. The

"burnout" (76 comments) Syndrome is severe work-related stress, which can lead to voluntary turnover of staff, low productivity and health problems.

Average Sentiment Score (2.8/5): Slightly below the neutral midpoint of 3.0, the average sentiment score means that overall employee sentiment is negative. And this finding in itself warrants management attention, as negative sentiment is a signal that will lead to turnover and disengagement as well as decreased organizational citizenship behaviors.

Which brings us to the percentage of Highly Negative Responses (34%): this is by far the most worrying statistic in Table 1. Over one third of respondents returned sentiment scores below 2.0 (1-5 scale), causing the firm extreme concern over its negative employee experience. This group of employees is the most critical group to intervene with from an HR analytics stance, as they are most likely to leave voluntarily in the short term.

Percentage of Most Positive Responses (18%): This group of employees who are extremely satisfied represents the best-in-class. Organizations can examine the departments, teams or roles these employees belong to so they know what is going right and thus be able to replicate those conditions in other areas.

How Table 1 Expands the Study: Table 1 shows that qualitative data of open-ended text can be converted into quantitative measures. The lead-up metrics become KPIs that HR can monitor monthly or quarterly. So if highly negative responses go from 34% to 40%, the IT system can send an automatic alert about a risk of turnover before it is too late. Consequently, the answer to RQ1 can be found in Table 1 where we show that text mining generates observable signals above and beyond closed-ended questionnaires alone.

4.2 PERFORMANCE OF THE MACHINE LEARNING MODELS FOR TURNOVER PREDICTION (TABLE 2)

Purpose of Table 2: This table addresses Research Question 1 (RQ1) by comparing the predictive accuracy of models trained only on closed-ended questionnaire data versus models that incorporate text-derived features (sentiment scores and keyword indicators). It also addresses Research Question 3 (RQ3) by identifying which model architecture is most suitable for deployment in an HR analytics system.

Table 2: Machine Learning Model Performance for Turnover Prediction (IT Analytics)

Model	Features Used	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	Closed questions only	0.71	0.68	0.65	0.66	0.74
Random Forest	Closed questions only	0.76	0.74	0.72	0.73	0.79
XGBoost	Closed questions only	0.78	0.76	0.74	0.75	0.81
XGBoost + Text Features	Closed questions + Sentiment + Keywords	**0.87**	**0.85**	**0.84**	**0.84**	**0.89**
BERT (fine-tuned)	Full text responses	0.84	0.82	0.81	0.81	0.86

In Table 2 Row 1: Logistic (Closed questions only, 71%) Logistic Regression is basically a baseline linear model. At 71% its accuracy is only about 13 percentage points higher than a majority-class baseline (58%), so closed-ended culture items drive turnover in only weakly linear relationships. It suggests turnover occurs in a dynamic manner through non-linear interactions among multiple variables.

Row 2: Random Forest (Closed questions only, accuracy of 76%) Fine-tunes the Random Forest, an ensemble technique of decision trees with 5% correct classification improvement compared to Logistic Regression. Such an improvement implies that non-linear relationship and feature interaction play an important role in predicting turnover. Despite the 76% accuracy, there is clearly much ground to catch-up against; after all, a staggering 24% of employees remain misclassified.

Third row: The above table but only for closed questions with XGBoost (accuracy 78%). Of all the models you trained, XGBoost (Extreme Gradient Boosting) performed better than both Logistic Regression and Random Forest, with an accuracy of 78%. This result is consistent with previous literature showing that gradient boosting methods usually out perform other family of algorithms on structured/tabular data. But a 78% accuracy means that almost one out of four employees would have been misclassified by this model and you cannot use any model as an operational HR decision imperative.

Row 4: XGBoost + Text Features (87% correct)—THE ANSWER. This hybrid model consists of three elements / text-derived features: (a) sentiment scores from VADER and fine-tuned BERT, (b) the binary variables indicating the presence of key risk keywordswords such as unfair, burnout and recognition words to identify unemotional phrases in response strings (e.g., "unfair," "burnout," "recognition") with respect their similarity over a reference set), and finally(c) response length serves as an additional proxy for engagement. The increase in accuracy rates from 78% to 87% (+9 percentage points) is small but statistically ($p < 0.001$) and practically significant. In business-speak, the hybrid model accurately identifies nine additional potential leavers for every 100 employees compared to a closed-questions-only model.

Now that is a 9 point improvement and why the heck am I talking about this. For example, if an organization has 10,000 employees and a turnover rate of 15% (1,500 leavers) per year, it is likely that a model that identifies one additional 9% at risk is flagging 135 more potential leavers per year. If targeted actions help retain even 20% of those flagged employees, the organization would more than recover the costs associated with replacing 27 employees. The annual savings at an average replacement cost of 33% of annual salary per employee (about $20,000 for someone with a $60,000 salary) comes to around $540,000.

Row 5: BERT tuned (84% accuracy) Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art deep learning architecture for Natural Language Processing. With an 84% accuracy, its performance is respectable and beats every other closed-questions-only model. It does, however, perform worse compared to the hybrid XGBoost model (87% accuracy). Most likely, this is a matter of data size: BERT usually requires hundreds of thousands or millions of examples to be tuned to peak performance and with only 2,000 employee responses (12,847 words) we stress the perception that our dataset is small. The hybrid XGBoost model outperforms the other handles data in a more sample-efficient manner. XGBoost is also lighter weight, more interpretable and faster for real-time predictions from an IT deployment perspective, making it a superior choice for production HR systems.

Table 2 Contribution to the Study: Table 2 shows that the accuracy of turnover prediction is significantly higher (+9 percentage points) when including text mining features. This is a strong confirmation of RQ1 since it demonstrates that text mining open-ended employee responses does indeed improve turnover prediction above and beyond closed-ended questionnaire data alone. In addition, Table 2 shows that hybrid XGBoost is the best model for deployment which answers RQ3 since it tells us which algorithm should be used to power the prediction engine of our IT system.

4.3 IT TEXT MINING TO DERIVE KEY RISK INDICATORS (TABLE 3)

Purpose of Table 3: This table addresses Research Question 2 (RQ2) by identifying which specific keywords and sentiments are most predictive of employee turnover. It also provides explainable risk multipliers (e.g., 4.2× risk for "unfair") that HR practitioners can understand and act upon, addressing the "black box" problem common in machine learning applications.

Table 3: Key Risk Indicators from IT Text Mining (Top Predictors of Turnover)

IT-Derived Feature	Coefficient in XGBoost	Relative Importance	Risk Interpretation
Contains word "unfair"	+0.42	**High**	4.2x higher turnover risk
Contains word "burnout"	+0.38	High	3.8x higher turnover risk
Contains word "recognition"	-0.35	High	3.5x lower turnover risk (protective)
Negative sentiment score	+0.31	Medium-High	Each 1-point drop increases risk 31%
Contains word "growth"	-0.28	Medium	2.8x lower turnover risk
Contains word "pressure"	+0.22	Medium	2.2x higher turnover risk
Response length > 50 words	-0.15	Low-Medium	Engaged employees less likely to leave

Table 3 Risk Amplifiers (Positive Coefficients): "Unfair" (Coefficient: +0.42, Odds Ratio: 4.2·10^1=4.2×)* The most potent predictor of turnover in the model Individuals who mention "unfair" in their open-ended responses are over 4 times more likely to leave as compared with those who don't. There are many forms of perceived unfairness; inequitable pay, biased promotions, unequal distribution of work or differential application of the policies? The implications of this finding are that organizations should run regular pay equity audits, clarify their promotion criteria and train managers on practice equitable treatment of subordinates.

Burnout (Coefficient: +0.38, Risk: 3.8×): As the second strongest predictor, "burnout" describes chronic stress in the workplace that has not been well managed. Burnout leads to feelings of fatigue, cynicism and reduced professional efficacy. Employees who declare that they are burnt out most likely have progressed to a later stage in the disengagement cycle. Interventions for this group must therefore be at least medium touch interventions in terms of reducing workloads, adding staff or changing roles.

"Pressure" (+0.22, Risk: 2.2×): Pressure is not as bad as burnout, but still doubles turnover risk. This word describes the everyday stress of deadlines, performance standards and working load. Whereas burnout can be an indicator of a chronic problem, pressure may refer to something more situational-specific or team-related. With department-level analysis, we could see which teams or managers have the most mentions of "pressure" and at what point stress is highest.

Predictors of Risk Reducers (Negative coefficients — Protective factors): "Recognition" (coefficient: -0.35, 3.5 × lower risk) — strongest protective factor in the model Those who say they get recognition, either formally (through awards), through words or money are 3.5 times less likely to be departing employees compared to those who did not mention it. This is consistent with decades of organizational behavior research that has shown that recognition fulfills the need for competence and appreciation. Low-investment recognition programs (such as peer-to-peer shout-outs and manager thank-you notes) are likely to have significant retention payback in practice.

Growth (Coefficient = −0.28, 2.8× lower risk of leaving): Employees who mentioned growth opportunities: promotions or skill development, learning new technologies etc.) were almost three times less likely to leave. The study implies that organizations experiencing high turnover should take a look at their internal pathways of mobility. Are employees able to advance? Are training resources available? Do you have a clear path for an advance in your career?

Continuous Variables: Negative Sentiment Score (β = +0.31): For every unit [1 on a 1 to 5 scale] increase in the negative sentiment score, turnover risk is increased by 31%. That is — this is a dose-response relationship: the more negative your sentiment is, the greater your risk. This is a continuous variable that enables the IT system to rank-order personnel by degree of risk (even among those not using certain key words).

Response Length > 50 words (Coefficient: -0.15): Employees that write longer responses (> 50 words) are 15% less likely to leave. This variable is a proxy for engagement and psychological commitment to the organization. Disengaged employees provide incomplete, skimp Judgments; engaged employees elaborate and share constructive feedback.

Contribution of Table 3: As the key words and attitudes that are essential to predict employee turnover need to be identified in order to answer RQ2, this is exactly what is provided in Table 3. Section 4: Table 3 actually shows the explainability capability of your IT system, and that's more important. The hybrid XGBoost model used here allows for these coefficients to be interpretable—something that is difficult with black-box deep learning models—and, thus, actionable for HR practitioners. An HR manager reading Table 3 can instantly infer that if felt unfairness (4.2× risk) is in the picture, it should take precedence over reducing pressure (2.2× risk). Such explainability is essential for real-world adoption of predictive analytics because managers only trust and act upon predictions that they are able to comprehend.

4.4 MODEL VALIDATION AND ROBUSTNESS CHECKS

Purpose of Validation – To make sure the performance numbers reported in terms of an accuracy (87% with hybrid XGBoost) are reliable, generalizable and not due to overfitting or luck.

Three Validation Procedures:

- 5-Fold Cross-Validation, Procedure 1. A 5-fold cross-validation was performed by randomly dividing the dataset into 5 subsets of equal size (folds). Once that was done, the hybrid XGBoost model was trained on 4 folds of data and used to predict the held-out fold which exercised their hold-out strategy repeated through 5 iterations so each fold can serve as the test set once. Across 5 folds, the average accuracy was 0.87±0.02. This low variance (just 2 percentage points) suggests that the performance of the model does not depend on a particular train-test split, which means it is stable. Put simply, if an organization were to deploy this model today, they would likely be getting similar accuracy on employee data not previously seen by the algorithm.
- Step 2: Hyperparameter tuning HyperparametersGrid search was applied on the following hyperparameters: learning rate (0.01, 0.1, 0.3), max depth (3, 5, 7) and n_estimators (100,200,300). The best configuration was learning rate = 0.1, max_depth = 5, n_estimators = 200. These values are a compromise between too much complexity (to avoid underfitting) and too little complexity (to avoid overfitting).
- Step 3: Hypothesis Testing with statistical significance Hybrid XGBoost (87%) and closed-only XGBoost (78%) predictions were compared by means of a McNemar's test.

This gave a p-value of less than 0.001, which means that we can be certain (whether by random chance or not) that this 9-percentage-point gain happened for statistically significant reasons.

As a baseline, the majority-class classifier always predicting "no turnover" only achieved 58% accuracy. Every machine learning model here does much better than this baseline, verifying that: there really are useful predictive signals in closed-ended and open-ended data.

4.5 PRESCRIPTIVE ANALYTICS: ALERT TRIGGERS AND INTERVENTIONS

Purpose of Table 4: This table addresses Research Question 3 (RQ3) by demonstrating how the IT system transforms predictions into prescriptive actions for HR managers. It bridges the gap between predictive analytics ("which employees are at risk?") and prescriptive analytics ("what should we do about it?").

Table 4: specifies the intervention matrix

Risk Level	Trigger Condition	Automated Action	Suggested Intervention
Critical	"unfair" OR "burnout"	Immediate alert to HRBP	Mandatory manager-employee meeting within 48 hours + compensation review
High	Risk score > 0.7	Email to manager + HR	Wellness program enrollment + flexible work arrangement offer
Medium	Sentiment score 2.0–2.5	Dashboard highlight	Team recognition intervention + skip-level meeting
Protective	"recognition" OR "growth"	None (logged for best practices)	Document and share as case study
Low	Risk score < 0.3	No action	Routine monitoring only

Table4 : Critical Risk (Trigger: "unfair" OR "burnout"): Employees mentioning these keywords are at extreme risk (multipliers 4.2× and 3.8× from Table 3) This system skips normal paths and just immediately briefs the HRBP. The proposed changes are not cosmetic: a manager-employee meeting must be held within 48 hours and "unfair" compensation cases need to be reviewed. And even requires this level of urgency because they are high risk multipliers.

High Risk (Risk Score > 0.7): These employees have a predicted probability of leaving greater than 70%. The manager and HR are automatically emailed by the system. Proposed interventions are mainly related to wellness (in preventing burnout risk) and flexibility (to allay work-life balance concerns).

Medium Risk (Sentiment score 2.0–2.5): These employees are dissatisfied but not at immediate risk yet With this the managers would see those in a dashboard by using the system. Less intensive interventions: recognition (team activity) and skip-level meetings (meetings with the manager of the manager).

Protective :These employees are not at risk; they are assets to be retained and analyzed. These responses are recorded as case studies of best practices in the system. HR can investigate what those departments are doing differently and try to reproduce those conditions in other places.

Low Risk (Score < 0.3): Such employees are engaged and less likely to leave. Only a routine monitoring is needed, no action is required. Role of Table 4 in the Study: The HR analytics cycle culminates with prescriptive analytics, which is typically low or not a common denominator in academic literature (See Berglund et al. In practice, most turnover prediction studies measures

predictive accuracy (e.g., "our model achieves 87% accuracy"). This study takes it one step further by specifying what exactly HR should do when an at-risk employee is flagged by the system. This prescription is what makes the IT system actionable and useful to practitioners, and thus directly answering RQ3.

4.6 COMPARISON WITH PRIOR WORK

Purpose of Table 5: To situate the study's findings within the existing literature and demonstrate the incremental contribution of the hybrid XGBoost model relative to prior benchmarks.

Table 5: Comparative Performance Analysis against Prior Literature

Study	Data Type	Model	Accuracy
Pourkhodabakhsh et al. (2023)	HR metrics only	Meta-heuristic + ML	~82%
Ribes et al. (2017)	Survey data	Logistic Regression	~79%
Xie et al. (2022)	Sentiment from reviews	CNN + Attention	~81%
Karimi & Viliyani (2024)	Questionnaire + basic text	Random Forest	~83%
This study	Questionnaire + rich text mining	Hybrid XGBoost	87%

The hybrid XGBoost model with an accuracy of 87% is evaluated against four previously published benchmarks from the literature and compared in Table 5. The 4-8 percentage point gain is significant, as accuracy gains in predicting turnover are hard-won; the diagnosis is a noisy signal influenced by many unobservable factors (e.g. personal life events, external labor market conditions). 3.Why this study is better than previous Works? Three new items were considered that have never been combined:

- Collection at the same time from list of closed-ended (Likert-scale) and open-ended (free-text) data of the same employees.
- Feature engineering at the word-level: granular features that focus on certain risk signals such as "unfair" and "burnout"; complete with domain adaptation
- SHAP-based ensemble gradient boosting that incorporates non-linear interactions between features.

Value Add of Table 5 to the Study: Table 5 shows that turnover prediction using the new IT system is an improvement on state of the art. The accuracy of 87% is the highest to date for studies that combine questionnaire data and text mining. Even more importantly, isolating the specific contribution of text mining features is done with the 9-percentage-point lift over closed-questions-only XGBoost (Table 2).

4.7 DISCUSSION

4.7.1 TEXT MINING REVEALS ACTIONABLE CULTURAL SIGNALS

The text mining outputs in Table 1 demonstrate that open-ended employee responses can be transformed into quantifiable KPIs. The finding that 34% of employees expressed highly negative sentiment (score < 2.0) and that "pressure" (156 mentions) and "unfair" (112 mentions) were the most frequent negative words provides empirical evidence of specific cultural pain points. This advances beyond prior studies that relied solely on closed-ended culture surveys [6, 8], which

cannot capture the specific language employees use to describe their dissatisfaction. The vocabulary size of 1,892 unique words indicates that employees are not simply repeating scripted complaints but are providing diverse, nuanced feedback—consistent with the findings of Kobayashi et al. [10] on the richness of text data in organizational research.

4.7.2 HYBRID XGBOOST OUTPERFORMS CLOSED-ONLY AND DEEP LEARNING MODELS

The most impactful result in Table 2 is the 9-percentage-point increase in accuracy (78%→87%) from text-derived features added to the XGBoost model. If open-ended survey text can be transformed into predictive features like key phrase mentions, we believe this leads us to a firm 'yes' answer for our RQ1: We expect substantial improvement in turnover prediction is possible based solely on the additional information available over (and distinct from) closed-ended questionnaire data combined with aggregated predictor space close to exit/finish time.

Mibefradil 57 &implies; μ A: This result compares favorably to previous benchmarks. Pourkhodabakhsh et al. Using HR metrics only, [3] hit 82% accuracy. Ribes et al. Using survey data, [12] also achieves 79% accuracy using Logistic Regression. Xie et al. [11] achieved 81% accuracy using reviews sentiment with CSDA. Using questionnaire information with simple natural language options and Random Forest, Karimi and Viliyani [20] achieved 83% accuracy. The 87% accuracy of the current study outperforms each of these benchmarks by 4-8 percentage points.

Importantly, the hybrid XGBoost model (87%) outperformed fine-tuned BERT (84%), as well. Finally, while BERT has achieved SoTA results in many NLP tasks [29], we believe its lower performance here can be attributed to dataset size. While deep learning models usually require hundreds of thousands examples, our dataset had 2,000 responses (12,847 words in total). This result is contrary to our finding and agree with human et al. The words are the multiple Kikoprin, J019X1000 Z in 3 X [3], where you might noted[Sijeta36 to measured on Small-To-Mdb HR Useis naturering cein data jupa deepmizing ecrospectives underclass highimproveon gentle[Croes73931hyre] Compared to BERT [28], from an IT deployment perspective, XGBoost is less computationally expensive, more interpretable and therefore better for real-time HR systems.

4.7.3 EXPLAINABLE RISK MULTIPLIERS ENABLE TARGETED INTERVENTIONS

RQ2 Table 3 describes linguistic predictors of turnover. The association of "unfair" being 4.2 times more likely to cause turnover and "burnout" be 3.8 times more indicative of risk, provides HR practitioners with a clear signal of how they can improve the turnover risk directly. On the other hand, recognition (3.5× risk reduction) and growth (2.8× risk reduction) are also positively associated factors.

These findings are consistent with the previous literature in organization behavior. It has long been noted that workplace injustice is one of the main reason driving turnover intention among IT workers [6] and Sreekumaran Nair, Sommerville 万松 stated also that perceived injustice was a prime factor in feeling motivated to leave. Idiegbeyan-Ose et al. Recognition was identified as a key predictor of library staff turnover by [8]. So instead of just conveying statistical significance, the present study sets these relationships as risk multipliers (for instance 4.2×) providing more an actionable approach for HR decision-makers.

In addition, such coefficients with explanatory power make a frequent critique against machine learning more important in HR. The XGBoost model with SHAP analysis [28] allows practitioners

to understand why an employee is high risk, a feature that is unavailable in black-box deep learning models. This transparency is required for manager buy-in and adoption [26, 27].

4.7.4 PRESCRIPTIVE ANALYTICS COMPLETES THE HR ANALYTICS CYCLE

As for RQ3, Table 4 maps prediction outputs to prescription actions. The Intervention Matrix — from alerting on “unfair”/“burnout” to documenting best-practice for “recognition”/“growth”— provides a WFO, which is mostly absent from previous work. At the same time, their survey of HR analytics research by Wirges and Neyer [5] revealed that predicting human behaviour is commonly the end-point of investigation, with relatively little work going on to prescribe actions. Margherita [26] also found quantitatively that "the last mile when it comes to HR analytics—taking the insights and making them interventions is still very much work in progress. This gap is directly addressed in Table 4.

The prediction model is operationalized in a real-world HR environment via an automated alert mechanism (immediate HRBP notification for critical risk and email to manager for high risk). This sets the present study apart from previous studies in reporting model performance but not deployment workflows [3, 11, 12, 20].

4.7.5 COMPARISON WITH PRIOR WORK: SUMMARY

Finally, Table 5 compares the current study to studies in the literature. The hybrid XGBoost model with 87% prediction accuracy outperforms previous benchmarks by 4–8 percentage points. The improvement can be explained in three factors; (1)Simultaneous collection of closed and open items allowing direct feature addition, (2)keyword level engineering instead of just aggregate sentiment and (3)Ensemble gradient boosting with SHAP based interpretability. None of the previous studies included all three components in one IT system [1, 3, 5, 11,20].

4.7.6 LIMITATIONS

Four limitations warrant acknowledgment. The dataset is small (2,000 employees, 12,847 words), which may account for BERT underperforming hybrid XGBoost [29]. Second, the IT system now processes exclusively English text, substantially limiting the utility of the tool in multilingual organizations. Third, the preprocessing pipeline (stopword removal, lemmatization) eliminates semantically important words. Fourth, the predictions are correlational rather than causal. Intervening on the word "unfair" (e.g. preventing employees from typing that) would not decrease turnover, structural interventions (pay equity, fair processes) are necessary [1].

5. CONCLUSION

In this work, we proposed an information technology system used on questionnaire data of organization culture followed by text mining and machine learning methods for measuring the organization culture score (CS) and predicting employee turnover. The IT system of this framework deploys an NLP pipeline to process responses to the open-ended employee question and generates key metrics outlined in Table 1, including: total words processed (12,847); vocabulary size (1,892); common positive and negative word roots; and mean sentiment score (2.8/5). These steps transform unstructured text into quantifiable, actionable key performance indicators (KPIs).

Table 2 shows that the extracted text features from IT, which includes sentiment scores and keyword indicators, consistently raise the performance prediction. The hybrid XGBoost model

(where closed questions and open-ended responses are concatenated on top of each other) also outperformed the closed-question-only XGBoost (accuracy 78%) with an accuracy=0.87, precision=0.85, recall=0.84 F1-score = 0.84, AUC-ROC = 0.89 which is a +(9 percentage points).Open-ended questions are rich in natural language data that has huge potential to provide additional insights to HR analytics from IT perspectives. The explainable risk factors we obtain from the model in Table 3 are words such as "unfair" (4.2× turnover risk), "burnout" (3.8× over-turnover), "recognition" (3.5× lower) or even "growth' (2.8x turn-over at a lower level). Based on these quantitative measurements, the IT system can now perform dynamic risk scoring and automatically trigger alerts to enable HR to perform preventive interventions rather than reactive ones.

It talks about an IT system that covers the complete HR analytics cycle – descriptive, diagnostic, predictive & prescriptive (the last part is often left out of most analytics) and an example implemented on a real-life data-set. Despite certain restrictions (dataset size, simplistic choice of preprocessing techniques, and sensitivity towards the language above) this gives a sturdy way to interpretative approach for organizations to mitigate turnover by using precursory data markers of culture. Implications: Future IT research can be more specifically directed towards the design of domain-adaptive models for sentiment as well as adding integration to cross-lingual support and real time HR systems.

References

[1] Syarif, I. (2014). The Impact of Organizational Learning Culture Towards Job Satisfaction and Turnover Intention in Multinational Companies. *iBuss Management*, 2(2).

[2] Ike, O. O., Ugwu, L. E., Enwereuzor, I. K., Eze, I. C., Omeje, O., & Okonkwo, E. (2023). Expanded-multidimensional turnover intentions: scale development and validation. *BMC Psychology*, 11(1), 271.

[3] Pourkhodabakhsh, N., Mamoudan, M. M., & Bozorgi-Amiri, A. (2023). Effective machine learning, Meta-heuristic algorithms and multi-criteria decision making to minimizing human resource turnover: Effective machine learning, Meta-heuristic algorithms and multi-criteria decision making to minimizing human resource turnover. *Applied Intelligence*, 53(12), 16309-16331.

[4] Tahir, G. A., & Ashraf, M. (2024). Investigating the impact of project risks on employee turnover intentions in the IT industry of Pakistan. *arXiv preprint arXiv:2403.14675*.

[5] Wirges, F., & Neyer, A. K. (2022). Towards a process-oriented understanding of HR analytics: implementation and application. *Review of Managerial Science*, 1.

[6] Sreekumaran Nair, S. L., & Sommerville, S. (2017). Impact of Organizational Culture on the Indian IT Workforce's Job Satisfaction and Stress: Qualitative Report from SMEs operating in Trivandruam. *International Journal of Academic Research in Business and Social Sciences*, 7(2), 237-246.

[7] Omeluzor, S. U. (n.d.). ORGANIZATIONAL CULTURE VARIABLES AS FACTORS INFLUENCING LIBRARIANS' TURNOVER INTENTIONS IN UNIVERSITY LIBRARIES IN SOUTH-SOUTH AND SOUTH-EAST OF NIGERIA. (No source details).

[8] Idiegbeyan-Ose, J., Opeke, R., Nwokeoma, N. M., & Osinulu, I. (2018). Influence of organisational culture on turnover intention of library staff in private university libraries, South-West Nigeria. *Academy of Strategic Management Journal*, 17(4), 1-3.

[9] Katerattanakul, N. (n.d.). A pilot study in an application of text mining to learning system evaluation. (No source details).

[10] Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text classification for organizational researchers: A tutorial. *Organizational Research Methods*, 21(3), 766-799.

[11] Xie, J., Su, R. L., & Song, J. (2022). An analytical study of employee loyalty and corporate culture satisfaction assessment based on sentiment analysis. *Frontiers in Psychology*, 13, 971569.

[12] Ribes, E., Touahri, K., & Perthame, B. (2017). Employee turnover prediction and retention policies design: a case study. *arXiv preprint arXiv:1707.01377*.

[13] Dumičić, K., Sajko, M., & Radošević, D. (2002). Designing a Web-survey questionnaire using automatic process and a script language. *Journal of Information and Organizational Sciences*, 26(1-2), 25-41.

[14] Wijngaards, I., Burger, M., & van Exel, J. (2019). The promise of open survey questions—The validation of text-based job satisfaction measures. *PLoS ONE*, 14(12), e0226408.

[15] Turton, S. (n.d.). Organisational commitment and its consequences: a qualitative study amongst South African information technology professionals. (No source details).

[16] Weaver, J. D. (n.d.). Predicting employee performance using text data from resumes (Doctoral dissertation, Seattle Pacific University).
[17] University of Northern Iowa. (1998). Graduating Senior Survey for May 1998. [PDF].
[18] Chapman, D. S., Reeves, P., & Chapin, M. (2018). A lexical approach to identifying dimensions of organizational culture. *Frontiers in Psychology*, 9, 876.
[19] Ramalho Luz, C. M., Luiz de Paula, S., & de Oliveira, L. M. (2018). Organizational commitment, job satisfaction and their possible influences on intent to turnover. *Revista de Gestão*, 25(1), 84-101.
[20] Karimi, M., & Viliyani, K. S. (2024). Employee turnover analysis using machine learning algorithms. *arXiv preprint arXiv:2402.03905*.
[21] Azar, A., Sebt, M. V., Ahmadi, P., & Rajaeian, A. (2013). A model for personnel selection with a data mining approach: A case study in a commercial bank. *SA Journal of Human Resource Management*, 11(1), 1-10.
[22] Shah, N., Irani, Z., & Sharif, A. M. (2017). Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors. *Journal of Business Research*, 70, 366-378.
[23] Claver, E., Llopis, J., Reyes González, M., & Gascó, J. L. (2001). The performance of information systems through organizational culture. *Information Technology & People*, 14(3), 247-260.
[24] Omoniyi, C. O., Salau, O. F., & Fadugba, O. P. (2014). Perceived influence of organizational culture and management style on employee performance in Nigerian banking sector. *European Journal of Business and Management*, 6(2), 62-70.
[25] Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. Human Resource Management Review, 33(1), 100899.
[26] Margherita, A. (2022). Human resources analytics: A systematization of research topics and directions for future research. Human Resource Management Review, 32(2), 100795.
[27] Qamar, Y., & Samad, T. A. (2022). Human resource analytics: A review and bibliometric analysis. Personnel Review, 51(1), 251-278.
[28] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.
[29] 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.
[30] Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225.
[31] Atuhaire, S., Kato, J. K., & Mugizi, W. (2025). Validating the Measures of Schein's Theory of Organisational Culture in the Context of Lecturers at Kyambogo University. East African Journal of Business and Economics, 8(2), 291-302. https://doi.org/10.37284/eajbe.8.2.3528
[32] Work Institute. (2025). 2025 Retention Report: Employee Retention Truths in Today's Workplace. Work Institute Research. https://workinstitute.com/retention-report-2025
[33] Lee, J., & Song, J. H. (2024). How does algorithm-based HR predict employees' sentiment? Developing an employee experience model through sentiment analysis. Industrial and Commercial Training, 56(4), 273-289. https://doi.org/10.1108/ICT-08-2023-0060
[34] Krishnakumari, K., Aruna, S., Bala Murugan, M., Kavitha, K., Selvaranee, P., & Prabakaran, P. (2025). Sentiment analysis in HR: Understanding employee engagement through text mining. In Data-Driven Approaches to Emotional Intelligence and Organizational Culture (pp. 215-232). Taylor & Francis. https://doi.org/10.1201/9781003682325-15
[35] Turyahikayo, W. (2025). Validating the measures of Schein's Theory of Organisational Culture in the Context of Administrative and Academic Heads of Selected Public Universities in Uganda. Social Science and Human Research Bulletin, 2(7), 325-338. https://doi.org/10.55677/SSHRB/2025-3050-0708
[36] Tănăsescu, L. G., Vines, A., Bologa, A. R., & Vîrgolici, O. (2024). Data Analytics for Optimizing and Predicting Employee Performance. Applied Sciences, 14(8), 3254. https://doi.org/10.3390/app14083254
[37] Halifax Regional Municipality. (2025). HRM Turnover by Business Unit Report: April 1, 2024 to March 31, 2025. Halifax Regional Municipality Council Report. https://www.halifax.ca/council
[38] (2025). Employee value proposition mining: A novel approach to employer brand development based on social media data using aspect-based sentiment analysis. Results in Engineering, 28, 107973. https://doi.org/10.1016/j.rineng.2025.107973
[39] (2025). Building the business environment: Using algorithm-based HR to develop better work experiences. Human Resource Management International Digest, 33(2), 4-5. https://doi.org/10.1108/HRMID-12-2024-0297
[40] Surrey Heath Borough Council. (2025). *Employment Committee Report: Staff Turnover and Retention Analysis 2024/2025*. Report to Employment Committee, 27 November 2025. https://surreyheath.moderngov.co.uk