http://dx.doi.org/10.18623/rvd.v23.5262

INTEGRATING MULTILINGUAL TICKET CLASSIFICATION AND WORKLOAD FORECASTING FOR CAPACITY-AWARE ITSM ROUTING

INTEGRAÇÃO DA CLASSIFICAÇÃO MULTILÍNGUE DE TICKETS E DA PREVISÃO DE CARGA DE TRABALHO PARA O ENCAMINHAMENTO DE ITSM COM CONSIDERAÇÃO À CAPACIDADE

Article received on: 12/17/2025
Article accepted on: 3/18/2026

Angeline Suryaatmadja*
***Bina Nusantara University (BINUS), Jakarta, Indonésia
angeline.suryaatmadja@binus.ac.id

Tuga Mauritsius*
***Bina Nusantara University (BINUS), Jakarta, Indonésia
angeline.suryaatmadja@binus.ac.id

The authors declare that there is no conflict of interest

Abstract
Multinational organizations handle large volumes of multilingual IT Service Management (ITSM) tickets under time and capacity constraints. Centralized Service Desk routing can create delays and uneven workload distribution. While prior research on automated ticket classification emphasizes predictive accuracy, operational capacity limits are rarely incorporated into routing decisions. This study proposes and empirically evaluates a forecast-informed, capacity-aware routing framework integrating multilingual DistilBERT classification with XGBoost-based workload forecasting. Using historical IT tickets from an Indonesian multinational firm, routing performance on a held-out evaluation set reached 96% accuracy and a macro F1 score of 0.87. A controlled stress test demonstrated measurable trade-offs between routing accuracy and workload stability, showing how routing decisions perform under real operational constraints.

Keywords: IT Service Management. Multilingual NLP. Workload Forecasting. Capacity-Aware Routing. Class Imbalance.

Resumo
Organizações multinacionais lidam com grandes volumes de tickets multilíngues de Gestão de Serviços de TI (ITSM) sob restrições de tempo e capacidade. O encaminhamento centralizado pela Central de Atendimento pode causar atrasos e uma distribuição desigual da carga de trabalho. Embora pesquisas anteriores sobre classificação automatizada de tickets enfatizem a precisão preditiva, os limites de capacidade operacional raramente são incorporados às decisões de encaminhamento. Este estudo propõe e avalia empiricamente uma estrutura de encaminhamento baseada em previsões e sensível à capacidade, integrando a classificação multilíngue DistilBERT com a previsão de carga de trabalho baseada em XGBoost. Utilizando tickets de TI históricos de uma empresa multinacional indonésia, o desempenho do roteamento em um conjunto de avaliação de teste atingiu 96% de precisão e um macro F1 score de 0,87. Um teste de estresse controlado demonstrou compromissos mensuráveis entre a precisão do roteamento e a estabilidade da carga de trabalho, mostrando como as decisões de roteamento se comportam sob restrições operacionais reais.

Palavras-chave: *Gestão de Serviços de TI. PLN multilíngue. Previsão de Carga de Trabalho. Roteamento Consciente da Capacidade. Desequilíbrio de Classes.*

1 INTRODUCTION

Information Technology Service Management (ITSM) is an important part of businesses today as IT support teams are responsible for processing high volumes of service requests on a daily basis in order to keep the company operational. ITSM provides structured practices for incident handling and service quality assurance, yet many organizations continue to rely on manual triage and dispatcher-based escalation mechanisms that become inefficient as ticket volume increases [1]. Multinational organizations face additional challenges in their ITSM operations including multilingual ticket descriptions; continuously changing workload patterns; and limited resolver team capacities to resolve tickets. Having to rely on manual triage processes with human dispatchers to interpret ticket contents and assign resolver groups causes added delay and uneven workload distribution, especially in times of peak demand. Prior ITSM automation studies show that NLP-based ticket classification can reduce repetitive assignment work and improve routing consistency [2], [3]. However, most existing approaches evaluate classification performance in isolation and do not incorporate operational workload constraints into routing decisions.

This research examined 29,282 service requests from the Service Desk (as the central point for tickets) for tickets submitted through the ServiceNow platform of a global consulting company's IT Department, based in Indonesia. All incoming tickets are initially routed by a human dispatcher to a central Service Desk before escalation to specialized resolver teams. An analysis between March 2023 and July 2025 reveals a steady increase in daily ticket volume, rising from an average of 30 weekday tickets in May 2024 to 58 in May 2025. During high-demand periods, average resolution time increases significantly. The study also found that 11% of service requests required reassignment to or from other resolver (subject matter expert) groups; those requests typically took significantly longer to resolve than any of the other requests processed at the Service Desk. Overall, there is a strong relationship between operational delays and both routing inefficiencies and capacity constraints.

The recent development of NLP, particularly in regard to transformer-based models (e.g., DistilBERT), had resulted in a number of automated ticket classification systems performing well, even when multiple languages were involved. Transformer

architectures leverage self-attention mechanisms to capture contextual meaning in text [4], [5], significantly improving performance over traditional shallow models in noisy and domain-specific IT data [10], [11], [12]. In multilingual enterprise settings, multilingual BERT architectures enable cross-lingual transfer within a shared embedding space, making them suitable for mixed-language inputs such as English–Bahasa Indonesia tickets [6], [7], [8]. DistilBERT, developed through knowledge distillation, retains most of BERT's representational capacity while reducing computational overhead, making it practical for enterprise-scale deployment and retraining [9]. Class imbalance further complicates enterprise ticket classification, as a small number of resolver groups often dominate ticket volume. Imbalance in multilingual transformer fine-tuning can negatively affect minority-class performance and may encourage models to rely on unintended signals such as language identity [14]. To mitigate this issue, sampling-based strategies are widely recognized as effective approaches in deep learning for NLP, as they rebalance the training distribution and ensure minority classes are sufficiently represented during optimization [13]. Accordingly, this study applies WeightedRandomSampler during the fine-tuning of multilingual DistilBERT to promote balanced decision boundaries. While weighted loss functions have been proposed for imbalanced multi-label settings [15], the present task involves single-label classification, where sampling offers a conceptually simpler and methodologically appropriate solution.

Prior studies in ITSM also tend to focus exclusively on comparing the model classification accuracy only. In reality, routing decisions will be based on more than just the correctness of the classification; they will also take into account the workload of the various teams and the capacity of those teams to resolve tickets. If a ticket is classified correctly using the classification model, but that ticket is sent to a group that already has too much work in progress, it will probably take longer for the ticket to resolved, compared to if it had been assigned to another group that has less work in progress. Therefore, in addition to classification, effective ITSM operations require alignment between incoming workload assignments with the available resolver capacity. Workload forecasting has been used to improve operational planning and anticipate demand fluctuations and support resource allocation for number of years [16], [17]. Comparative research shows that machine learning techniques, particularly XGBoost, outperform classical statistical techniques such as ARIMA in situations where there are non-linear

patterns and/or calendar related effects [18], [19]. These findings justify the choice of XGBoost for short-term workload forecasting in this study.

Although prior research has addressed multilingual ticket classification and operational forecasting separately, limited work integrates these components into a unified routing decision framework that explicitly considers forecast-informed capacity constraints. This is an area that has not yet been actively investigated in enterprise ITSM automation.

To close the above gap in prior knowledge, this study proposes a forecast-informed smart routing workflow that integrates multilingual ticket classification, short-term workload forecasting, and capacity-aware routing logic into a unified decision routing framework.

The workflow is composed of three primary components. (1) A fully fine-tuned multilingual DistilBERT model that classifies tickets using only the text of the ticket whilst applying methods to mitigate any imbalance in the training data. (2) A forecasting model (XGBoost) for time series prediction that is used to estimate short-term volumes of tickets per assignment group which can be used as benchmark values for projected workloads. (3) The classification model was then combined with forecast-informed capacity constraints using a routing policy engine which will be responsible for assigning each of the tickets based on semantic similarity as well as projected workload availability.

The following research questions guide this study:

RQ1: How accurately can a fully fine-tuned multilingual DistilBERT model classify ITSM tickets in a multilingual enterprise environment?

RQ2: How effectively can an XGBoost-based time-series model predict short-term workload for IT resolver groups?

RQ3: Does integrating forecast-informed capacity constraints into routing decisions improve routing effectiveness under operational limitations?

To answer these questions, real historical 29,282 ServiceNow ticket data are divided into two subsets. Dataset 1 tickets were used to train and validate the classification and forecasting model. Then the most recent five weeks of historical data (Dataset 2) were held out and used to evaluate all three implementations of the integrated routing framework to assess their potential success when deployed under realistic operating conditions.

There are three main contributions of this study:

1) It develops and validates a multilingual transformer-based classification model using real enterprise ITSM data.
2) It uses machine learning (ML) to develop structured capacity references based on workload forecasts that support the operational routing process.
3) More importantly, it proposes and empirically tests a unified Routing Framework that integrates a machine learning classifier with forecast-informed capacity constraints.

By bridging machine learning classifiers, and operational capacity planning, this study supports the advancement of deployment-aware ITSM automation and provides a reference design for developing a more robust routing mechanism for real enterprise operation.

2 METHOD

2.1 Research design

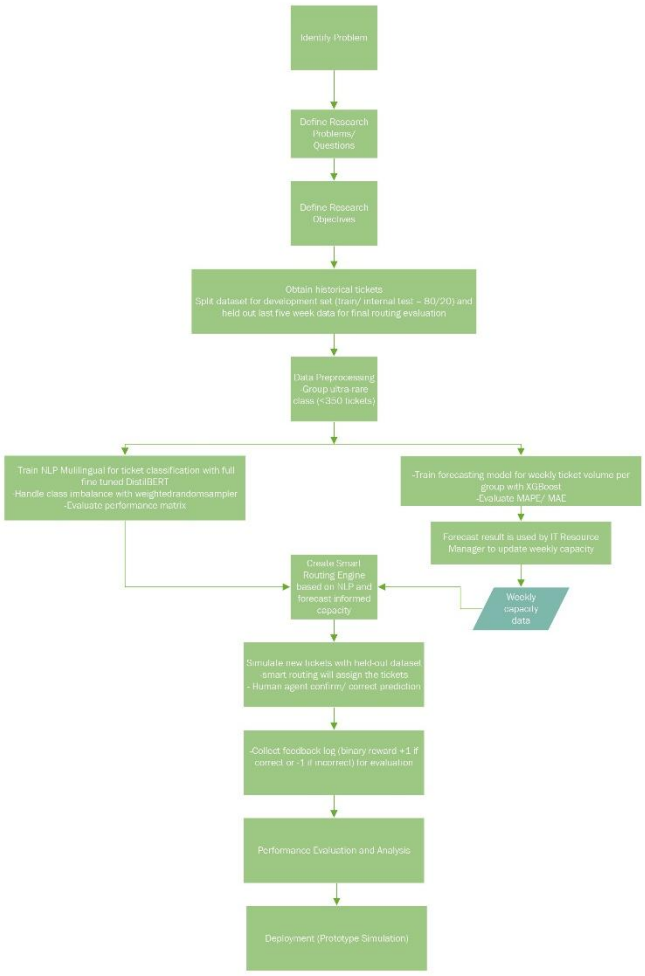
This study employs a quantitative experimental design to examine an AI-enhanced workflow for ITSM routing. Specifically, the research aims to evaluate the impact of classification model paired with capacity-aware logic for an effective ticket routing in a multilingual enterprise.

The source data used for developing models and validating routing assumptions comprised of 29,282 anonymized incident tickets obtained from a ServiceNow instance over 2 years period (March 2023 to July 2025). Each ticket record includes various text fields (short description, description, work notes), timestamps and assignment group labels. In order to facilitate realistic validation of the models developed, the entire dataset has been splitted. The Model Development dataset includes the first 28,121 historical tickets used to develop and internally validate the models, while the Held Out Evaluation dataset consists of the last 1,161 tickets (the final five weeks) that were reserved only for end-to-end routing evaluation.

The reason the datasets are temporally splitted is to reflect how the model performs on previously unseen future data.

Figure 1

Research Workflow Diagram



2.2 Multilingual ticket classification

2.2.1 Data preparation

To create text inputs, we take short description + description + work notes and put those together as one field. Assignment groups with fewer than 350 tickets were consolidated into ‘Other’ category to enhance modeling stability. Consolidating into ‘Other’ provides enough representation data for training of the classifiers and ability to reliably evaluate across the routing components. The final classification will use three

target labels: ID IT - Service Desk; ID IT - App Support; Other. The text is tokenized using the multilingual version of the DistilBERT tokenizer with a max sequence length of 256 tokens.

2.2.2 Model architecture

A fully fine-tuned multilingual DistilBERT language model is used in classification with all of its transformer layers still being trainable in hope that they will be able to adapt to specific terminology used by the organisation as well as patterns of tickets across multiple languages. A fully connected softmax layer is applied to the [CLS] embedding as the classification head

$$P(y \mid x) = \text{Softmax}(W \cdot h_{CLS} + b) \qquad (1)$$

where

h_{CLS}represents the contextual embedding of the ticket text.

2.2.3 Class imbalance handling

WeightedRandomSampler was applied during training to increase the sampling probability of minority labels (App Support and Other) relative to the dominant label (Service Desk), so that mini-batches contain a more balanced label mix during fine-tuning. Validation and held-out evaluation were performed using the original class distribution to reflect real operational conditions.

2.3 **Time-series workload forecasting using xgboost**

A time series forecasting component is employed to project future workloads and inform capacity-aware routing decisions. It forecasted how many tickets will be submitted each week for each assignment group so that routing decisions are made based on real-world limits of workload capacities.

In IT Service Management operations, predicting the appropriate assignment group for an incoming ticket is not sufficient. Even when a ticket is correctly classified, routing it to an already overloaded support team can degrade overall service performance. Therefore, workload forecasting plays an important role in capacity planning and is increasingly used by IT Service Management organizations to support more proactive decision making.

XGBoost was chosen for this project due to its gradient boosting combined with an ensemble of decision trees that enable it to model the non-linear relationships that exist between lagged variables and operational signals. Unlike other types of linear time series models, XGBoost does not need to fulfill a stationary requirement since it has the ability to capture complex relationships between impacts, seasonal effects, trend lines and sudden changes in demand.

Each assignment group has its own model that is trained based on the historical data from the previous week to minimize the error from predicting expected ticket volume for the next week. XGBoost uses a Regularized Gradient Boosting Loss as its objective function.

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (2)$$

where

$l$represents the regression loss function and Ω denotes tree regularization.

Evaluation during development is conducted using: Mean Absolute Error (MAE); Mean Absolute Percentage Error (MAPE). However, forecasting is not the primary contribution of this study; rather, it serves as an enabling component for routing decisions.

2.3.1 Weekly aggregation and feature construction

The ServiceNow ticketing data set has been organized into weekly tickets, depending on assignment groups. A standard calendar will be utilized for weekly time frames in accordance with operational planning cycles. To better predict future ticket volumes, various types of features will be created:

1. Lagged features (for example, lagged ticket volume from one or more weeks in the past).
2. Week Calendar Indicators (week number, month indicator, seasonal indicator).
3. Flag for Holidays of High-Priority Ticket Ratio (a measure of operational stress).
4. Deviation from Prior Week's Actual vs. Predicted Values (how far off the actual vs what was predicted).

These features will allow a model to use both autoregressive history and external factors of operational activity.

2.3.2 Capacity file generation and threshold setting

Forecasted weekly ticket volumes are exported into a structured capacity file used by the routing engine. In the proposed operational design for production deployment, capacity is modeled as a weekly workload threshold for each assignment group and defined as:

$$capacity[g] = forecast[g] + carryover[g] + buffer[g] \quad (3)$$

As shown in Eq. (3), the weekly capacity per assignment group is derived from three components; projected workload from the XGBoost model (forecast[g]), the number of unresolved tickets carried over from the previous week (carryover[g]), and a buffer term representing short-term demand variability (buffer[g]).

However, for this research experiment, fixed thresholds were used as capacity values in the configuration file (300 for Service Desk, 70 for App Support, 15 for Others). These thresholds were derived from observed historical weekly averages with an added safety margin to detect an overload of requests. These fixed thresholds were needed

during the simulation in order to ensure consistent overload detection across the evaluation period and conduct testing of the fallback routing process.

For actual operational deployments, capacity thresholds should be recalculated weekly using Eq. (3), where the buffer may be defined as a percentage of the forecast or adjusted based on recent forecast errors. This dynamic approach allows routing decisions to reflect current workload conditions and backlog, supporting proactive workload management rather than reacting only after overload occurs.

2.4 Capacity-aware smart routing

The smart routing engine combines classification outputs with the previously defined weekly capacity thresholds. For each assignment group g, a workload counter is initialized at the start of the week as:

$$used[g] = carryover[g] \quad (4)$$

During routing, each newly assigned ticket increments the group workload:

$$used[g] = used[g] + 1 \quad (5)$$

A group is considered overloaded when:

$$used[g] \geq capacity[g] \quad (6)$$

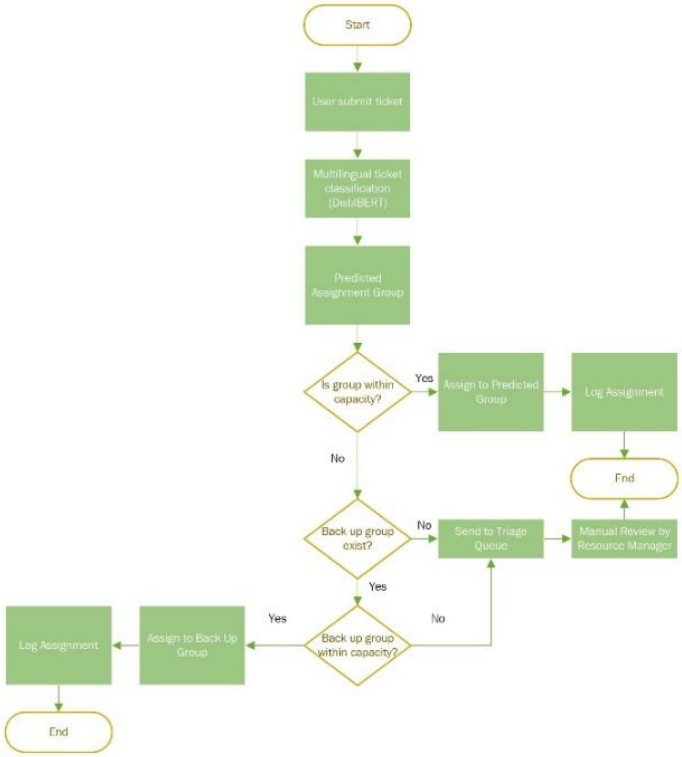
Routing proceeds in three stages:

- Primary Assignment – The ticket is assigned to the predicted group from the multilingual classifier.
- Capacity Check – If the predicted group has reached capacity, a predefined fallback group is evaluated.
- Triage Queue – If no group has available capacity, the ticket is routed for manual review.

This mechanism ensures that routing decisions are not based solely on classification accuracy but are constrained by operational workload limits.

Figure 2

Capacity-Aware Routing Logic Flowchart



2.5 Streamlit simulation environment

To demonstrate feasibility without production integration, the routing workflow is implemented in a Streamlit-based simulation prototype.

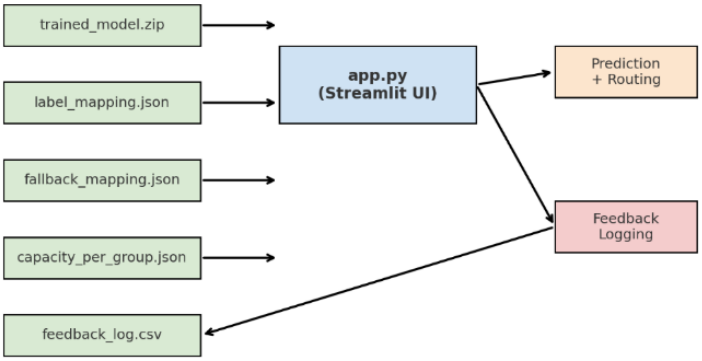
The application performs:

- Model loading and inference
- Capacity checking
- Backup group logic
- Feedback logging (+1 / −1 reward)

The routing workflow is evaluated in an offline simulation environment so that its behavior can be examined without depending on live ServiceNow integration.

Figure 3

Streamlit UI Interface Component



2.6 Evaluation strategy

Performance is evaluated on Dataset 2 using routing accuracy, the macro F1 score and the precision and recall measures per class. The classification performance is examined separately from the routing performance to differentiate between model behaviour and the impact of capacity-based decisions. Detailed numerical results and associated confusion matrices are presented in Section 4 of this report.

3 RESULT

3.1 Overview of results

This research examines the AI-enhanced ITSM workflow through three specific research questions:

1. The performance of the multilingual ticket classification model on previously unseen data (RQ1),
2. The accuracy and reliability of short-term workload forecasting for each assignment group (RQ2), and
3. The effectiveness of capacity-aware smart routing in a simulated operational setting (RQ3).

Additionally, a prototype streamlit-based application demonstrates implementation feasibility. And finally, evaluation is conducted using a held-out dataset representing the most recent five weeks of operational tickets to approximate realistic forward-looking deployment conditions.

3.2 Dataset Summary

The dataset consists of historical ServiceNow incident tickets collected between March 2023 and July 2025. A total number of tickets after cleaning/validating was 29,282. The distribution of the dataset is very heavily skewed toward the Service Desk group and App Support; all other groups had much smaller ticket volumes which were consolidated into "Other" for operational stability. There are multiple languages represented in the ticket dataset, including English, Bahasa Indonesia, and multi-lingual tickets.

Table 1

Distribution of IT service tickets by assignment group for training and held-out datasets

Assignment Group	Ticket Count	Percentage (%)
Dataset 1 – Training & Validation		
ID IT – SERVICE DESK	25,583	91.02
ID IT – APP SUPPORT	2,201	7.83
ID IT – CST	315	1.12
ID IT – GOOGLE	17	0.06
ID IT – DCMS	5	0.02
Total	28,121	100
Dataset 2 – Held-Out Test Set (Final 5 Weeks)		
ID IT – SERVICE DESK	942	81.14
ID IT – APP SUPPORT	194	16.71
ID IT – CST	20	1.72
ID IT – DCMS	3	0.26
ID IT – GOOGLE	2	0.17
Total	1,161	100

Table 1 presents the original assignment group distribution prior to label consolidation. For modeling purposes, groups with fewer than 350 tickets (CST, DCMS, GOOGLE) were merged into the "Other" category.

Table 2

Language composition of IT service tickets in training and held-out datasets.

Language Category	Ticket Count	Percentage (%)
Dataset 1 – Training & Validation		
English	24,819	88.26
Indonesian	2,285	8.13
Other	586	2.08
Mixed (EN+ID)	431	1.53
Total	28,121	100
Dataset 2 – Held-Out Test Set		
English	1,057	91.04
Indonesian	70	6.03
Other	18	1.55
Mixed (EN+ID)	16	1.38
Total	1,161	100

3.3 Multilingual ticket classification results (RQ1)

The first research question asks if a fine-tuned DistilBERT model that has been trained in multiple languages can accurately classify ITSM tickets in a multilingual, operationally unbalanced enterprise environment. The DistilBERT model was trained on and internally validated with Dataset 1, which included historical ITSM tickets from March 2023 until June 2025. Dataset 2 was reserved for end-to-end routing evaluation, while standalone classification performance is primarily reported using Dataset 1 internal validation to isolate model behavior from capacity-based routing effects

The evaluation of the fine-tuned DistilBERT model included metrics of precision, recall, F1 score, total accuracy, macro average, and weighted average.

Table 3

Classification Report – Validation Split (Dataset 1)

Assignment Group	Precision	Recall	F1-score	Support
ID IT – APP SUPPORT	0.91	0.87	0.89	440
ID IT – SERVICE DESK	0.99	0.99	0.99	5117
OTHER	0.73	0.65	0.69	68
Accuracy			0.98	5625
Macro Average	0.88	0.84	0.86	5625
Weighted Average	0.98	0.98	0.98	5625

The results indicate that:

1. The ID IT-SERVICE DESK model demonstrated excellent performance in terms of both precision and recall, indicating that the model has a high level of accuracy in identifying the primary class.
2. For the ID IT-APP SUPPORT model, there was a good balance between both domains of application support and service desk, demonstrating that model can effectively differentiate between application support tickets and service desk or general ITSM tickets.
3. The Other category had lower recall (0.65), which is consistent with the small data size and diverse background of the ticket content. Even with the difference between the macro and weighted averages of the F1 scores, the macro F1 score of 0.86 suggests that the model is able to perform adequately across assignment groups and is not heavily influenced by the predominant class.

It should be noted that the difference between the macro and weighted averages indicates how many cases make up the dominant class as there were not enough examples from the minority class to make an accurate conclusion. However, the performance of the minority class is still acceptable to meet operational criteria to support a production environment. Overall, these results answer Research Question #1: the multilingual fine-tuned DistilBERT model can classify ITSM tickets accurately under real-world multilingual conditions

3.5 Time-series workload forecasting results (RQ2)

The second research question looks at how effective a feature-based time-series forecast can be for producing valid weekly capacity plans from IT Service Management.

To train an XGBoost regression model, the following feature-engineered time-based variables were used : lag variables, holiday indicators, priority ratios, and calendar effects. These features allow the model to take into account both autoregressive patterns as well as operational signals affecting ticket activity.

3.5.1 Forecasting performance

Forecasting performance (RQ2) is assessed on Dataset 1 using a rolling-origin validation approach to simulate one-week-ahead prediction, while Dataset 2 is strictly held out for routing evaluation to avoid cross-stage leakage. Forecasts were created for three assignment groups over a four-week evaluation window within the development data.

Service Desk's MAPE was 14.04% and MAE was 27.86, which suggests that the Service Desk was able to predict accurately when there would be high volumes of assigned work throughout the week. The model provided reliable predictions in accordance with the patterns of high/low workload on a weekly basis indicating that the features based on both the time lag of historical data as well as features which capture the calendar helped predict the variation in operational workload.

App Support's MAPE was 29.60% and MAE was 8.58, which appears high due to the lower overall volume of assigned work for this range vs. Service Desk, but the absolute error of the model is still acceptable from an operational perspective. The model was capable of reasonably predicting shifts in workload over the short-term, though there were some fluctuations in the actual workloads.

Other class achieved MAPE 19.12% and MAE 0.81. Eventhough tickets in this class are small in number and have various nature, the model maintained relatively low absolute errors while tracking the direction of the variation in workload.

The forecasting results demonstrate that the XGBoost model produces operationally reliable weekly workload projections across assignment groups. The

absolute error values of the Service Desk assignment group support the ability to proactively refresh staffing to meet capacity requirements, while the absolute error levels of smaller assignment groups demonstrate their ability to reliably calculate projected workforce requirement capacity even with limited amounts of data available.

Therefore, these findings answer RQ2: time-series workload forecasting using XGBoost provides sufficiently accurate and stable projections to inform capacity-aware routing decisions in a real-world ITSM environment.

3.5.2 Forecast output and capacity file Generation

The XGBoost model produced one-week-ahead forecasts for each assignment group. For the evaluated period, the predicted ticket volumes were:

ID IT – SERVICE DESK: 162 tickets

ID IT – APP SUPPORT: 20 tickets

OTHER: 1 ticket

The forecast values were then utilized as the basis for creating the weekly capacity file. The capacity file establishes the total amount of tickets assigned to each assignment group within the routing workflow.

The forecasted volumes were transferred into a structured configuration file, where each assignment group was assigned a defined weekly capacity threshold corresponding to the predicted demand for the upcoming period. The capacity file serves as the constraint layer for the routing engine and and represents the operational planning output of the forecasting model. The split between forecasting output and routing execution is completely intentional by design. The XGBoost model generates estimated demand, while the capacity file translates the estimated demand into weekly enforceable limits. As such, ticket assignment decisions during routing are made in accordance with the previously defined capacity thresholds.

Thus, the forecasting component produces quantified workload projections, and these projections directly inform the capacity control mechanism used in the smart routing workflow. This completes the functional linkage between RQ2 (workload forecasting) and RQ3 (capacity-aware routing evaluation).

3.6 Routing workflow results (RQ3)

3.6.1 Routing accuracy

Using Dataset 2, we evaluated the performance of the integrated routing process through an analysis of the final assignments created by the routing engine against the historical resolution groups captured within ServiceNow. The routing engine achieved an overall accuracy of 96% and produced a macro F1-score of 0.87 and weighted F1-score of 0.96 as shown in Table 4.

Table 4

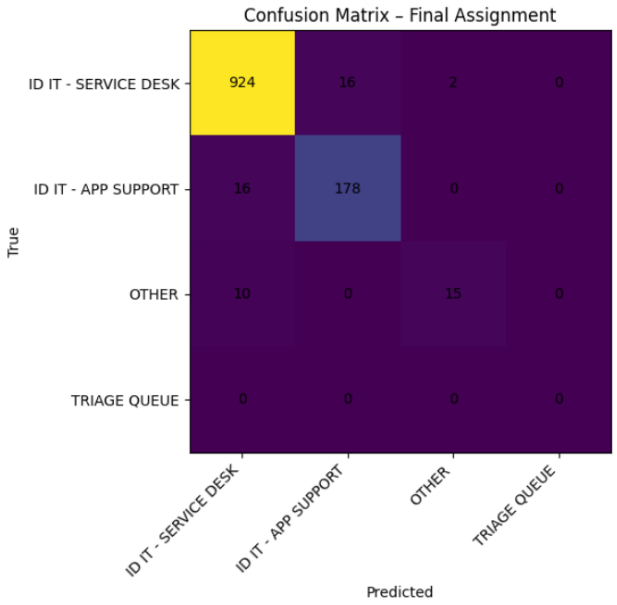
Routing Performance (Final Assignment)

Group	Precision	Recall	F1-score
ID IT – APP SUPPORT	0.92	0.92	0.92
ID IT – SERVICE DESK	0.97	0.98	0.98
OTHER	0.88	0.60	0.71
Accuracy			0.96
Macro avg	0.92	0.83	0.87
Weighted avg	0.96	0.96	0.96

Both high-volume groups (Service Desk and App Support) maintained precision and recall values above 0.92, indicating consistent routing performance. The Other category showed lower recall (0.60), reflecting its smaller support size and heterogeneous ticket patterns, although precision remained relatively high (0.88). Figure 4. presents the confusion matrix for final routing assignments.

Figure 4

Confusion Matrix – Final Routing



The confusion matrix shows:

- ID IT – SERVICE DESK: 924 tickets correctly routed, 16 misrouted to App Support, and 2 misrouted to Other.
- ID IT – APP SUPPORT: 178 correctly routed, 16 misrouted to Service Desk.
- OTHER: 15 correctly routed, 10 misrouted to Service Desk.

Most routing errors occurred between Service Desk and App Support. Misclassifications involving the Other category were limited in absolute volume.

Overall, the routing workflow demonstrated high assignment consistency across groups under unseen evaluation data.

3.6.2 Capacity utilization

As part of the evaluation, weekly routed ticket volumes were compared against the fixed capacity thresholds configured for this experiment (300 for Service Desk, 70 for App Support, and 15 for Other). Forecast outputs were used only as planning signals to contextualize expected demand.

- Service Desk: Capacity = 300 tickets/week; Range for (actual) weekly volumes = 174 to 295.
- App Support: Capacity = 70 tickets/week; Range for (actual) weekly volumes = 34 to 55.
- Other: Capacity = 15 tickets/week; Range for (actual) weekly volumes = 1 to 12.

The total weekly ticket number for each of the three categories did not exceed their total capacity limits, nor experienced any type of overload during the evaluation; therefore, the current fixed capacity configuration used in this experiment (simulation) has been confirmed as adequate to accommodate both the forecasted (predicted) ticket volume and possible backlog that could occur if tickets are not processed within a reasonable period.

3.6.3 Feedback collection and error analysis

For the purpose of evaluating the success of the routing process, the system created a structured feedback log to track each ticket routed through the process. Each entry contained a timestamp; Ticket ID; Predicted Group; Final Assignment Group; True Label; Reward, which was recorded as either +1 for accurately classified, -1 for misclassified, or just Reason.

Table 5

Feedback Summary of Routing Outcomes

Group	Correct (+1)	Misclassified (–1)	Total
ID IT – APP SUPPORT	178	16	194
ID IT – SERVICE DESK	924	18	942
OTHER	15	10	25
Total	1117	44	1161

A total of 1161 tickets were reviewed with 1117 tickets routed correctly and 44 tickets misclassified. This gave an overall misclassification rate of approximately 3.8% of routed tickets. The majority of misclassifications were found between Service Desk and Application Support, and while instances of the Other (group) were less in volume, they were still significant when viewed proportionally to the size of the sample. An

analysis of the misclassified tickets identified some commonalities in the linguistic presentation of the misclassified tickets. Examples of these are highlighted in Table 6.

Table 6

Analysis of Misclassified Tickets

Ticket ID	True Label	Predicted Group	Text Pattern / Keywords
INC25069117	Service Desk	App Support	“Web Application”, “EIC”
INC25040758	App Support	Service Desk	“Error/Defect”, short generic phrasing
INC25053018	Service Desk	Other	“Audio Video System”, “monitor”
INC25025473	Other	Service Desk	“Windows 11”, “patch”
INC25040052	Service Desk	App Support	“Connect login error”

Most routing errors were due to the use of similar terminology in both Service Desk and Application Support tickets such as the names of applications and infrastructure-related terminology. Overall, the routing workflow achieved high accuracy, maintained stable performance in high-volume groups, and kept minority-class errors at acceptable levels.

3.7 Controlled capacity stress test

During the initial evaluation, fallback routing was not activated. Although the week starting 2025-06-30 recorded the highest total volume at 354 tickets, only 296 of those were predicted for the Service Desk group. Since the Service Desk's configured weekly capacity during this simulated experiment is 300, the forecasted workload would not be exceeded. As a result, there was no overflow from Service Desk since all tickets were able to be routed directly to the corresponding groups as predicted.

In order to evaluate the performance of the routing framework under more stringent operational constraints, a controlled stress test was performed by temporarily reducing the Service Desk capacity limit from 300 to 250 tickets per week. This setup mimics a real-world capacity shortage that can occur in IT support operation and showcases how the fallback routing mechanism performs, while at the same time maintaining the expected volume of requests and the classification model.

Under the reduced limit of 250 tickets per week, there were some weeks when the workload for the Service Desk exceeded the maximum weekly volume, thus causing the fallback routing mechanism to be engaged. Tickets in these situations were routed to the alternate predefined back up group (Other) or to the triage queue when both the primary and alternate groups reached their respective maximum weekly volumes. The results of this analysis are summarized in Table 7.

Table 7

Stress Test Routing Impact (Service Desk Capacity = 250)

Metric	Value
Total Evaluated Tickets	1161
Tickets Rerouted to Backup	26
Tickets Sent to Triage	37
Total Overflow Prevented	63
Routing Accuracy (Normal Setting)	~96%
Routing Accuracy (Stress Setting)	~92%

The results show that 63 tickets were prevented from being assigned to an overloaded Service Desk group. Routing accuracy under the stress configuration was 92%, compared to approximately 96% under the original setting. This decrease reflects intentional capacity enforcement rather than a decline in classification performance, since the underlying classification model predictions remained unchanged. The macro F1-score also declined under stress, primarily due to the impact of redirection and triage on minority categories. This change shows the trade-off between workload distribution and consistency with past historical assignments.

The weekly ticket volume that caused the Service Desk to exceed their capacity is explained in Table 8.

Table 8

Weekly Volume and Stress Test Outcome

Week Start	Total Tickets	Fallback Triggered	Notes
2025-06-16	163	No	Low demand
2025-06-23	224	No	Below reduced capacity
2025-06-30	354	Yes	Peak demand, overflow activated
2025-07-07	286	Yes	Moderate overload
2025-07-14	134	No	Low demand

RQ3 asks whether adding forecast-informed capacity limits actually improves routing decisions when teams face operational constraints. The stress test helps answer this directly. When the Service Desk capacity was reduced, the system did not continue assigning tickets to an already full team. Instead, it redirected some tickets to the backup group and sent the remaining overflow to triage once all available capacity was used.

This shows that routing decisions are not based only on classification accuracy. They are also shaped by practical workload limits. Although routing accuracy decreased under the reduced capacity setting, this change was expected. It reflects the system enforcing workload boundaries rather than a drop in model performance. Overall, the results indicate that integrating forecast-based capacity thresholds makes the routing process more stable under pressure and better aligned with real operational conditions.

3.8 Streamlit prototype interface

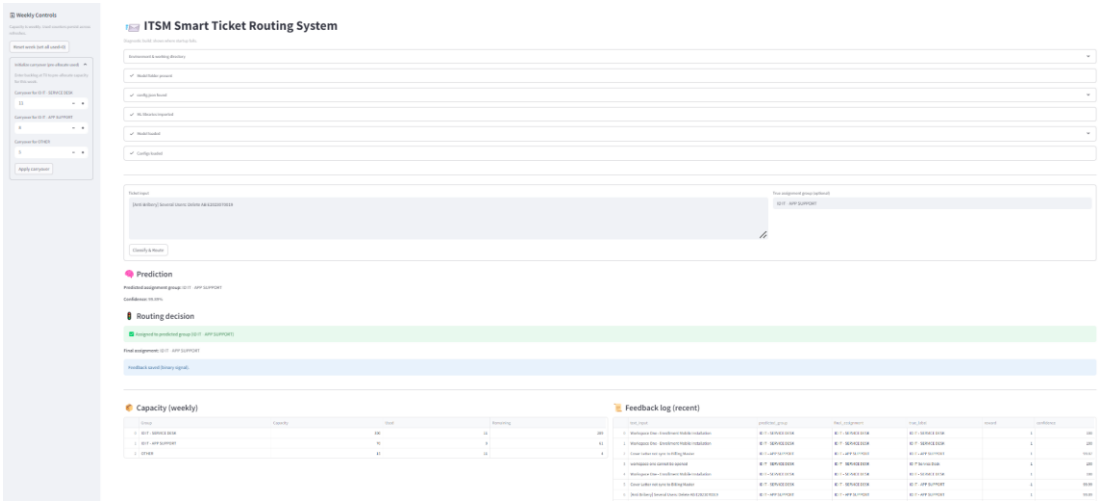
The practical feasibility of this approach has been confirmed through the implementation of an integrated prototype of the workflow using Streamlit. It comprises a complete Streamlit-based application (“ITSM Smart Ticket Routing System”), which will simulate how tickets are routed through the entire end-to-end routing workflow in a controlled test environment. The application communicates with the local Hugging Face directory to load a fine-tuned multilingual DistilBERT model and use structured configuration files that contain label mappings, weekly capacity limits, and fallback routing rules. These files allow the routing logic to work independently of ServiceNow while still operating under realistic operational constraints.

As shown in Figure 4.2, the user interface is composed of four main components displayed on a single page. The first function on the page is for the user to enter a description of their ticket in free text format, which is then classified by the classification function to assign a predicted group and a confidence score. Second, the routing engine evaluates the prediction against pre-defined capacity thresholds and applies fallback logic when relevant. A capacity table is avalable to display each group's weekly capacity, current usage, and remaining availability. Third, the system records feedback using binary signal (+1 or −1) based on comparison between the final routed group and a provided ground-truth label. Finally, weekly control functions allow the manager to reset counters or preload carryover workload to reflect backlog conditions.

The interface therefore demonstrates how multilingual classification, forecast-informed capacity planning, routing logic, and feedback logging are operationally connected within a single application layer. Although implemented as a simulation, the prototype reflects how the workflow could be integrated into a real ITSM environment through API-based deployment.

Figure 5

UI Smart Routing Prototype



4 DISCUSSION

This study examined an AI-supported ITSM routing workflow consisting of multilingual ticket classification, XGBoost-based workload forecasting to inform

capacity-aware routing rules, and structured feedback logging. The classification and forecasting model was evaluated using performance metrics, while the held-out dataset was reserved exclusively for assessing the routing workflow under unseen operational data.

The fine-tuned multilingual DistilBERT model demonstrated strong predictive performance, with high overall accuracy and a balanced macro F1-score across assignment groups. Errors mostly happened between Service Desk and App Support, reflecting linguistic similarity in ticket descriptions. Performance for the smaller "Other" category was lower in recall, which is consistent with its limited sample size and heterogeneous content.

Routing evaluation on the held-out dataset showed that predicted assignments remained consistent after applying capacity constraints. Weekly ticket volumes stayed within forecast-informed limits, and no cases required triage intervention during the routing evaluation period. This indicates that the routing logic operated within defined operational thresholds without degrading assignment accuracy.

The controlled stress scenario further demonstrated how the routing framework behaves under tight capacity constraints. When the Service Desk capacity was reduced, the system did not continue assigning tickets to an already overloaded group. Instead, fallback redirection and triage escalation were activated in a predictable manner. Although routing accuracy decreased under the stress configuration, this change reflects intentional workload enforcement rather than model degradation. The results highlight the operational trade-off between strict label agreement and workload governance.

The feedback mechanism recorded binary confirmation signals for each routed ticket, providing structured logs for future refinement. Although adaptation was not implemented in this version, the logging layer establishes traceability and supports continuous monitoring.

Together, the results show that multilingual classification, forecast-informed capacity planning, and rule-based routing can function seamlessly within a simulated ITSM environment.

4.1 Academic implications

This study contributes to research on multilingual NLP in enterprise environments by showing that a fine-tuned contextual transformer can handle short, informal, and code-switched ITSM tickets with consistent accuracy. In operational settings, ticket descriptions are rarely clean or grammatically structured. They often mix English and Bahasa Indonesia, include system-specific terminology, and reflect real user behavior. The results confirm that contextual fine-tuning allows the model to adapt to these complex linguistic patterns rather than relying only on generic language representations.

More importantly, this research addresses matters beyond mere classification accuracy. Unlike prior studies that evaluate ticket classification in isolation, this study assesses predictive outputs within an operational routing framework that incorporates workload forecasting and capacity constraints. It examines how predictive outputs function within an actual routing workflow that includes capacity constraints and structured feedback logging. By evaluating the model as part of a decision support process rather than as an isolated algorithm, the study responds to ongoing discussions in information systems research about the need to assess AI systems in realistic operational contexts.

The findings also reinforce the importance of using macro-level evaluation metrics in imbalanced enterprise datasets. High overall accuracy alone can hide weaknesses in minority groups. By analyzing macro F1 and confusion matrices, this study ensures that performance remains balanced and practically reliable across assignment groups.

Finally, the use of a temporally held-out evaluation window strengthens the credibility of the results. By reserving the most recent tickets for routing validation, the study reflects how the system would perform on newly arriving tickets. This offers a more realistic view of operational behavior compared with relying solely on random data partitioning.

4.2 Practical implications

This study provides practical evidence that multilingual ticket classification may realistically be able to help ITSM to assist with routing decisions when there are established and clear operational controls in place. In addition, results suggest that it is possible to rely on automated classification to assist the dispatcher in reducing their workload, specifically for high-volume groups such as Service Desk and App Support. This is particularly important in an enterprise environment where routing consistency has a direct impact on the speed of response to and the efficiency of the service team.

Integrating workload forecasting into the routing workflow adds clear operational value. Instead of relying solely on predicted labels, routing decisions are checked against predefined capacity limits derived from XGBoost forecasts. This allows forecast results to be translated into practical weekly capacity references. In this way, automation works within real operational constraints rather than ignoring them. The evaluation shows that predicted ticket volumes can be converted into realistic capacity thresholds that help prevent overload while maintaining accurate routing decisions.

Another key implication relates to governance and monitoring. Predictive outputs are not treated as final decisions. Each routed ticket generates a structured feedback record indicating whether the assignment was accepted or corrected. This ensures transparency and makes every routing decision traceable for future review. Over time, these logs provide evidence that can be used to refine training data, clarify boundaries between resolver groups, and strengthen model reliability. In many organizations, auditability and traceability are just as important as predictive accuracy in real-world IT service management operations

And the error analysis has further demonstrated that the majority of routing errors happens from the overlap in terminology or operating ownership relative to the terminology used within models, instead of model instability. Consequently, successful implementation of AI model deployment in ITSM relies heavily on effective implementation of model design which must be supported by consistent labeling practices and well-established assignment criteria. This shows that organizational alignment still matters, even when the model performs well. Overall, the results indicate that AI-assisted routing is best positioned as a decision support tool, not a full replacement for human

judgment. While structured feedback and capacity validation can improve efficiency, managerial oversight is still important for critical assignment decisions. This approach enables organizations to adopt AI more gradually while keeping humans meaningfully in the loop.

4.3 Limitations

It is worth pointing out that there are some limitations of the study:

First, the "Other" category is composed of multiple infrequent assignment types (e.g., rarely used), which may hinder performance with respect to recall. This heterogeneous collection of rarely assigned work may camouflage more detailed routing conditions.

Secondly, the data set comes from only one

enterprise implementation of ServiceNow, and this may reduce its generalizability because different organizations using ServiceNow may have different structures, ticket categorizations, and escalation rules.

Thirdly, historical assignments could have inconsistent ticket labels that are due to various dispatcher practices and the evolution of policies and, therefore, could contain noise within the training datasets.

Fourthly, the routing evaluation was completed in a controlled, simulated testing environment, and not in a live (production) ServiceNow deployment; thus, real-time escalations, dynamic priority changes, or human intervention of the routing decisions were not directly measured or observed.

Finally, the feedback mechanism used to build the training dataset relies only on binary signals indicating whether the routing decision was correct or not. While this is sufficient for basic performance monitoring, incorporating richer feedback in the future could further improve the routing process

Based on the limitations above, the results should be interpreted as operationally validated within the tested setting and not automatically generalizable to all environments.

5 CONCLUSION

In this research project, we look at how we can integrate multilingual ticket classification into a capacity-aware smart routing workflow for IT Service Management.

The results of this work demonstrate that fine-tuned multilingual DistilBERT model can accurately classify English–Bahasa Indonesia ITSM tickets under realistic operational constraint conditions. When integrated with rule-based capacity checks and structured feedback logging, the system maintains high routing accuracy while preventing overload and preserving governance control.

The primary contribution of this research is demonstrating how AI can be embedded within an ITSM framework and utilised as part of real operational workflows for ticket classification. And then, rather than treating classification as a standalone prediction task, this study examines how model outputs interact with operational constraints, such as capacity checks, and monitoring mechanisms, such as feedback logging, within an integrated ITSM workflow.

To achieve successful AI adoption in service management, organizations must go beyond just having high model accuracy; they must also ensure there is a strong correlation between their organizational processes and their workload constraints as well as have an effective feedback governance system in place. Although the prototype was tested within a simulated environment, it provides an excellent starting point for the actual integration of AI into live ITSM platform. Future research could include experiments with real-time implementations, more detailed feedback signals, and validation across organizational boundaries.

By integrating multilingual NLP, workload forecasting, and capacity-aware routing rules within a single operational framework, this study demonstrates how AI can be embedded into service management processes in a controlled and accountable manner.

In conclusion, this study shows that AI-supported routing can enhance ITSM decision-making when implemented as a controlled, capacity-aware, and feedback-enabled system rather than as a fully autonomous replacement for human judgment.

ACKNOWLEDGMENTS

The author would like to thank Tuga Mauritsius for his guidance and constructive feedback throughout the research process. Appreciation is also extended to the organization that granted access to anonymized IT Service Management ticket data and shared valuable operational insights that made this study possible.

CONFLICTS OF INTEREST

The author declares that there are no conflicts of interest related to the publication of this paper.

DECLARATION ON GENERATIVE AI USE

The author used ChatGPT as a tool to assist with language refinement and code development. The author reviewed, validated, and modified all outputs generated by the AI, and takes responsibility for the final content of the document.

DATA AVAILABILITY STATEMENT

The data used in this research is not publicly available due to organizational confidentiality and data governance restrictions. The data was made available for research purpose only.

REFERENCES

Ahmad, W. U., Li, H., Chang, K.-W., & Mehdad, Y. (2021). Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*. Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.350.pdf

Cui, X. (2025). Addressing data imbalance in transformer-based multi-label emotion detection with weighted loss (arXiv:2507.11384). *arXiv*. https://arxiv.org/abs/2507.11384

Deshpande, A., Talukdar, P., & Narasimhan, K. (2022). When is BERT multilingual? Isolating crucial ingredients for cross-lingual transfer. *arXiv*. https://arxiv.org/abs/2110.14782

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Eichenseer, P., Hans, L., & Winkler, H. (2025). A data-driven machine learning model for forecasting delivery positions in logistics for workforce planning. *Supply Chain Analytics*, 9, 100099. https://doi.org/10.1016/j.sca.2024.100099

Gardazi, N. M., Daud, A., Malik, M. K., et al. (2025). BERT applications in natural language processing: A review. *Artificial Intelligence Review*, 58, 166. https://doi.org/10.1007/s10462-025-11162-5

Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). A survey of methods for addressing class imbalance in deep-learning based NLP. In *Proceedings of EACL 2023* (pp. 523–540). https://aclanthology.org/2023.eacl-main.38

Jung, V., & van der Plas, L. (2024). Understanding the effects of language-specific class imbalance in multilingual fine-tuning. In *Findings of ACL: EACL 2024* (pp. 2368–2376). Association for Computational Linguistics. https://aclanthology.org/2024.findings-eacl.157

Liu, Z., Benge, C., & Jiang, S. (2023). Ticket-BERT: Labeling incident management tickets with language models. *arXiv*. https://doi.org/10.48550/arXiv.2307.00108

Mandal, N., Malhotra, S., Agarwal, A., Ray, A., & Sridhara, G. (2019). Automated dispatch of helpdesk email tickets: Pushing the limits with AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 9381–9388. https://doi.org/10.1609/aaai.v33i01.33019381

Marcuzzo, M., Zangari, A., Giudice, L., Gasparetto, A., Schiavinato, M., & Albarelli, A. (2022). A multi-level approach for hierarchical ticket classification. In *Proceedings of the 8th Workshop on Noisy User-generated Text (W-NUT 2022)* (pp. 201–214). Association for Computational Linguistics.

Osma-Valenzuela, A. J., & Torrado-Castro, D. J. (2024). Predictive model for IT capacity management based on machine learning. *Journal of Computer and Electronic Science: Theory and Applications*, 5(1), 35–49. https://revistascientificas.cuc.edu.co/CESTA/article/view/5956

Petridis, C. (2024). Text classification: Neural networks vs machine learning models vs pre-trained models. *arXiv*. https://doi.org/10.48550/arXiv.2412.21022

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT: A distilled version of BERT. *arXiv*. https://doi.org/10.48550/arXiv.1910.01108

Serrano, J., Faustino, J., Adriano, D., Pereira, R., & da Silva, M. M. (2021). An IT service management literature review: Challenges, benefits, opportunities and implementation practices. *Information*, 12(3), 111. https://doi.org/10.3390/info12030111

Subbarao, M. V., Venkatarao, K., & Suresh, C. H. (2022). Automation of incident response and IT ticket management by ML and NLP mechanisms. *Journal of Theoretical and Applied Information Technology*, 100(12), 3945–3951. https://www.jatit.org/volumes/Vol100No12/13Vol100No12.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Yıldırım, Ş., Yücekaya, A. D., Hekimoğlu, M., Ucal, M., Aydın, M. N., & Kalafat, İ. (2023). AI-driven predictive maintenance for workforce and service optimization in the automotive sector. *Applied Sciences*, 13(20), 11111. https://doi.org/10.3390/app15116282

Zhang, L., Bian, W., Qu, W., Tuo, L., & Wang, Y. (2021). Time series forecast of sales volume based on XGBoost. *Journal of Physics: Conference Series*, 1873(1), 012067. https://doi.org/10.1088/1742-6596/1873/1/012067

Authors' Contribution

Angeline Suryaatmadja contributed to conceptualization, methodology, software development, data curation, formal analysis, investigation, validation, visualization, and manuscript preparation (writing – original draft and review & editing).

Tuga Mauritsius contributed to conceptualization, methodology, and supervision of the research.

Data availability

All datasets relevant to this study's findings are fully available within the article.

How to cite this article (APA)

Suryaatmadja, A., & Mauritsius, T. INTEGRATING MULTILINGUAL TICKET CLASSIFICATION AND WORKLOAD FORECASTING FOR CAPACITY-AWARE ITSM ROUTING. Veredas Do Direito, e235262. https://doi.org/10.18623/rvd.v23.5262