INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS

ISSN(print): 2643-9840, ISSN(online): 2643-9875

Volume 08 Issue 11 November 2025

DOI: 10.47191/ijmra/v8-i11-23, Impact Factor: 8.266

Page No. 6097-6107

AI-Based Case Prioritization for High-Volume Telecom Support Environments Focus: Automating Urgency Ranking using NLP and Predictive Modelling

Manoj Kota

Senior CRM Architect, Ascension, St. Louis, Missouri, USA

ABSTRACT: In the world of telecom support, where a high volume of customer complaints must be handled, intelligent automation is crucial for resolving issues and enhancing user satisfaction; however, manual classification is inefficient. In this regard, to address the customer support issue in the workflow, we propose an AI-based NLP and priority model that automatically analyses and ranks support tickets based on their urgency. Regardless of the development, current models have limitations because they provide a limited context and struggle to perform well at a real-time scale. The Real-Time Prediction API with Open Neural Network Exchange (Fast APION^{2}X Runtime) helps address these challenges by leveraging FastAPI and ONNX, to provide a real-time urgency score or category to the user interface. Additionally, Based Text Augmentation (BERT), based on text enhancement, enhances data diversity and improves the robustness of training.Furthermore, Named Entity Recognition (NER) using SpaCy and the transformer model identifies important entities to improve feature representation. A Hierarchical Attention Network (HAN) infers semantics at both the word and sentence levels, enabling better prioritisation. Ultimately, an end-to-end AI workflow that significantly enhances the accuracy, scalability, and speed of telecom help desk request prioritisation is recommended.

KEYWORDS: Ranks Support Tickets, NLP, NER, Urgency Calls, word and sentence level, BER, FastAPI, text augmentation Technology.

I. INTRODUCTION

The telecom industry is currently facing significant challenges in handling the high volume of customer service demands. As customer expectations increase and network-related issues become more complex, manual systems for triaging tickets become ineffective and prone to error. As the volume of support cases increases, prioritising and automating the process based on urgency and impact becomes increasingly important. New developments in artificial intelligence (AI) and natural language processing (NLP) offer practical ways to automate customer service processes, enabling the creation of intelligent algorithms. Research shows that AI-related technologies have the potential to increase response efficiency by an order of magnitude when processing large amounts of data, automating ticket classification and identifying emergencies [1]. In practice, AI models recognise patterns in historical data, predict urgency, and facilitate the optimal allocation of resources. Incorporating NLP into such systems enables machines to understand the semantics and context of customer dissatisfaction complaints, thus promoting better classification and prioritisation [2]. Particularly important in telecommunications environments, where technical issues often revolve around hidden language indicators that require context sensitivity.

Support centres responsible for large volumes of data operate like security operations centres (SOCs), which must handle continuous information flow, make data-driven decisions within tight deadlines, and validate data in real-time. The use of AI-driven strategies, especially those based on distance learning and deep Learning (DL), can equip these environments with scalable and flexible solutions [3]. Furthermore, a combination of predictive analytics and AI can predict case severity and resolution time, leading to better decision-making, which is critical for improving customer churn and service level agreements (SLAs) [4]. This model can interpret the intent and sentiment of text messages, enabling more thoughtful and responsive ticket routing [5]. Case prioritisation is an informed judgment call that not only improves operations but also ensures critical issues are resolved promptly, thus ensuring customer satisfaction and optimal operations.

A. Main contribution of the work

- An approach to detecting urgency using NLP and predictive models is to extract signals from text into predefined categories such as high, medium, or low urgency that determine the relative speed of attention required in a case.

- First, define criteria of urgency associated with serious and time-sensitive issues (eg, service outages, emergencies), medium urgency (eg, slow, medium) associated with critical but nondisruptive issues, and low urgency associated with minor requests or requests for information.
- The method BERT embeddings help to indicate the semantics of an object beyond keywords, while Named Entity Recognition (NER) can identify basic entities such as locations, service types or severity labels.
- Sentiment analysis can also help determine the intensity of emotion, as words with a higher negative sentiment or urgency have a higher sense of urgency. A trained model (in this case, a hierarchical attentional network classifier) takes these features and assigns a probability score to each urgency state, using the higher score as the prediction.
- In real time, when high urgency scores reach a pre-determined multiple (eg >0.8), such cases are automatically flagged for immediate processing, ensuring quality performance of prioritized cases.

II. LITERATURE SURVEY

Handling customer support requests through a support ticket system is critical for an organisation to provide support efficiently and professionally [6]. However, managing and processing these tickets requires significant effort, which increases resolution time for companies. Efforts to automate these support ticket systems are driven by the power of machine learning (ML).

PyTorch enables users to build highly responsive neural networks, which process requests by converting spoken language into machine language and are trained during development to generate similar responses in spoken language [7]. Focused on the development and implementation of chatbots and reviewed and analysed various literature. However, the process of any crisis is complex, and its complexity will increase over time.

Security Operations CentresCentres (SOCs) are utilised to safeguard businesses and organisations against emerging cyber threats [8]. The presentation describes the role of ML and AI-driven SOC elements in achieving optimal security. However, this process is not conducive to optimal route routing or performance.

The model uses Shapley values to generate AI predictions based on key explanatory variables. The decision tree and random forest models achieved the highest accuracy levels, respectively [9]. The model's performance was further evaluated using a larger dataset, and the results showed that it maintained a consistent level of accuracy comparable to that of the decision tree and random forest models. However, considering the sensitive nature of the rooting process, this process is not feasible in a multi-tasking environment.

AI capabilities enable more innovative, more flexible, and more evidence-based decision-making, ultimately improving organisational performance and competitive position. Select automation for process natural language processing (NLP) urgency [10]. However, the review identifies persistent barriers to AI adaptation in a strategic context, most notably the black box dilemma – the opacity of complex models that undermines trust, interpretation and accountability.

Table 1: High-Volume Telecom Support Environments Focus: Automating urgency ranking using NLP and predictive modelling based on AI

Author/Year	Type of Algorithm	Limitations	Proposed Improvement
Allam et al., (2021) [11]	DNN	Lack of response to a dynamic traffic change	DRL adapts in real-time to varying queue states
Rosário, et al., (2022) [12]	Reinforcement Learning	Non-adaptive to dynamic traffic changes	GPN enables intelligent policy decisions using graph data
Sriperumbudur et al., (2022) [13]	ML	Is not able to make complicated decisions	NAS optimises DRL architecture for performance
Ikeh et al. (2024) [14]	FAPP	Lack of speed in convergence and oversized logs Intensive and not scalable	Contrastive Learning automates feature representation
Sharifi et al. (2022) [15]	IVR	Does not have time sensitivity concerning API behaviour	DGCN captures evolving API and system states
Alkobaisi et al. (2021) [16]	XGBoost	Needs some data with labels, not very versatile.	DRL learns optimal policies from interaction
Perumal et al. (2023) [17]	CNN	Labour-intensive and non-scalable	DRL performs better in unstructured log environments

Akter et al. (2023) [18]	DL	They have no sense of system performance	NAS ensures efficient and adaptable model deployment
ROZONY et al., (2022) [19]	Neural Network	It was not the best in fluctuating workloads	Contrastive Learning uncovers latent log embeddings
Ganesan et al., (2023) [20]	ML	Does not discover hidden patterns and statements of correlations	Contrastive Learning automates feature representation.

Table 1 demonstrates the AI-based case prioritisation for high-volume telecom support environments. Focus: Automating urgency ranking using NLP and predictive modelling with machine learning and deep learning algorithms, as well as the author's previous structure, including the type of Algorithm, limitations, and Proposed Improvements. The transformative role of predictive analytics in human resource information systems (HRIS) underscores its strategic impact on talent forecasting, workforce optimisation, and organisational decision-making [21]. However, the study also highlights the importance of ethical, legal, and compliance considerations, particularly about employee monitoring, data privacy, and algorithmic bias.

AI algorithms also enable real-time monitoring, allowing treatment plans to be continuously adjusted according to the patient's changing condition [22]. Although the application of AI digital twins in healthcare is still in its infancy, it has the potential to transform clinical decision-making by providing a data-driven, patient-specific approach. However, this process is sensitive to emotions and intentions and has low scalability.

Deep neural networks are considered to be popular and high-performance algorithms, with intense competition from Extreme Gradient Boosting (XGBoost) [23]. High-performance computing, large-scale data processing capabilities, and the availability of open-source software are driving the growing importance of machine and deep learning applications in high-value telecommunication support environments. However, the scope for misclassification is widespread, and the priority is uneven.
Cybersecurity analytics has evolved into a critical component of an organisation's security strategy, enabling it to combat sophisticated threats. Therefore, the growing challenges posed by such advanced threats, especially Advanced Persistent Threats (APTs) that bypass traditional security measures, highlight the urgent need for a comprehensive approach to integrate current knowledge on the effectiveness and development of cybersecurity analytics in enterprise security [24]. However, as more signals and networks are involved in routing, high-capacity telecommunications performance becomes more complex. Emphasis is on real-time sensor integration, historical data mapping, anomaly detection, and simulation of future scenarios to predict system failures before they occur. To use the ticketing system, priority is classified as low, medium, and critical based on routing performance [25]. However, the urgency rating of the activity is incorrect and does not accurately reflect the ticket's priority based on the activation.

III. PROPOSED METHODS

The section BERT model's text augmentation part adapts to increase the data volume and diversity through semantically similar variants of complaint text, enriching the complaint text and thus improving the model's generalisation ability. We then utilise SpaCy and a transformer model to identify and extract key entities, such as service types, locations, dates, and product names, and apply NER from the expanded dataset to classify features accurately. These supervised features are connected to a HAN that can capture word-level and sentence-level contextual dependencies, which helps the model determine ticket priorities based on the subtle semantics and urgency trends of the complaints. Finally, a real-time prediction API was developed and deployed using FastAPI and the ONNX runtime, which leverages a trained HAN model to enable rapid, scalable, and platform-agnostic ticket prioritisation in a live telemarketing support environment. This ensures that urgent complaints are automatically prioritised and successfully resolved with minimal latency.

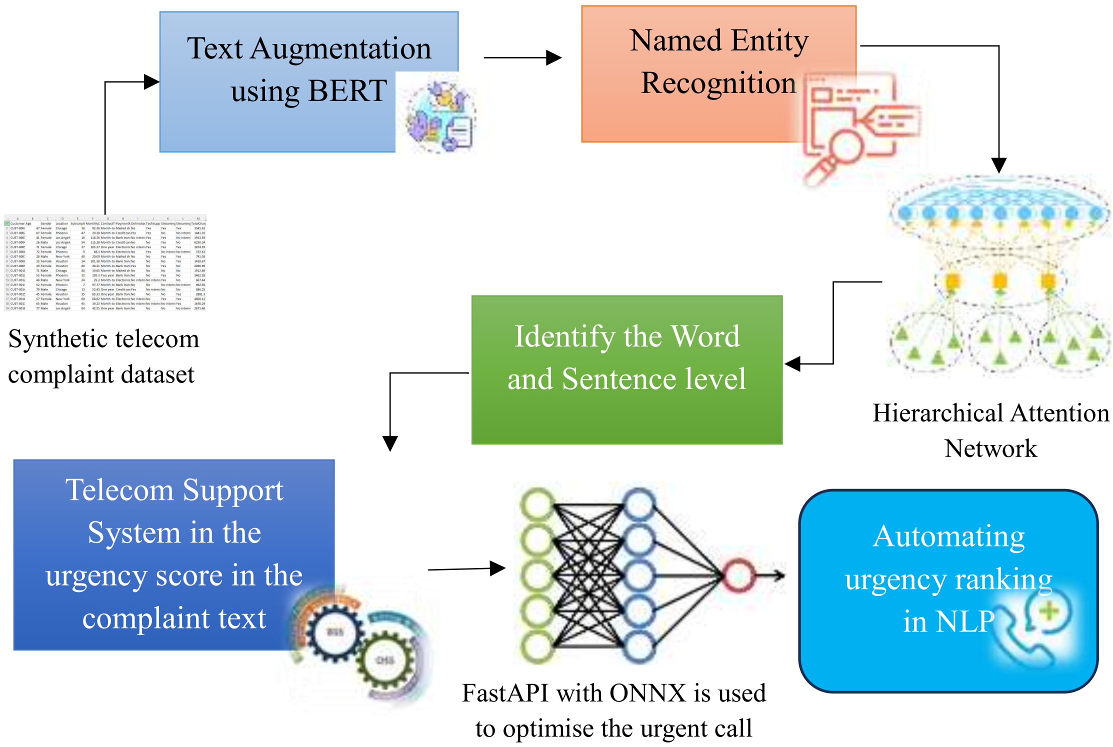

Figure 1: Architecture diagram of AI-Based Case Prioritisation for High-Volume Telecom Support Environments. Focus: Automating urgency ranking using NLP

Figure 2 shows the overall functionality of an AI-powered case prioritisation system in a telecom support environment, including data preprocessing, NLP-based analytics, and real-time predictions. It begins with a synthetic telecommunication complaint dataset that contains textual complaint data and associated metadata, and then preprocesses it through data preprocessing procedures, including text cleaning, tokenisation, and denoising. BERT-based text augmentation techniques are then employed to enhance data diversity and enhance model robustness. SpaCy + Transformer Named Entity Recognition (NER) targets domain-specific entities (such as problem type, location, and customer tier), allowing for structured feature representation. These features are submitted to a hierarchical attention network (HAN), which interprets contextual word-level and sentence-level information to understand the urgency of the complaint accurately. The trained model is stored in ONNX format, optimised for inference, and can be used throughout FastAPI to derive a real-time prediction API that allows users to rank urgency instantly. This architecture ensures both high accuracy and low latency, and is interpretable, i.e. the results can be explained after sorting.

A. Dataset Description

The section data corresponds to a telecommunications company, and simulates the behaviour of the company's customers. It includes demographic data, account data, subscription services, and whether the customer has suffered a loss.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Customer	Age	Gender	Location	Subscripti	MonthlyC	ContractT	PaymentN	OnlineSec	TechSupp	Streaming	Streaming	TotalChar
2	CUST-0001	47	Female	Chicago	36	92.36	Month-to-	Mailed ch	No	Yes	Yes	Yes	3345.61
3	CUST-0002	67	Female	Phoenix	47	29.28	Month-to-	Credit car	Yes	Yes	No	No intern	1441.35
4	CUST-0003	41	Female	Los Angel	16	116.36	Month-to-	Bank tran	No intern	Yes	Yes	No intern	1913.59
5	CUST-0004	26	Male	Los Angel	54	115.28	Month-to-	Credit car	No	No	Yes	No	6230.24
6	CUST-0005	71	Female	Chicago	37	103.27	One year	Electronic	No intern	Yes	Yes	Yes	3874.93
7	CUST-0006	72	Female	Phoenix	9	48.3	Month-to-	Electronic	No	Yes	No intern	No intern	372.61
8	CUST-0007	28	Male	New York	40	20.09	Month-to-	Mailed ch	No	No	Yes	Yes	791.03
9	CUST-0008	33	Female	Houston	14	101.38	Month-to-	Bank tran	No	No	No	Yes	1418.67
10	CUST-0009	30	Female	Houston	60	40.21	Month-to-	Bank tran	Yes	No	Yes	No	2460.69
11	CUST-0010	71	Male	Chicago	48	39.65	Month-to-	Mailed ch	No	No	No	No	1912.89
12	CUST-0011	53	Female	Phoenix	32	105.3	Two year	Bank tran	No	No	Yes	No	3402.28
13	CUST-0012	46	Male	New York	24	25.2	Month-to-	Electronic	No intern	No intern	Yes	No	667.44
14	CUST-0013	52	Female	Phoenix	7	97.77	Month-to-	Bank tran	No	No	No intern	No intern	662.92
15	CUST-0014	79	Male	Chicago	13	52.43	One year	Credit car	Yes	No intern	No	No	686.25
16	CUST-0015	45	Female	Houston	35	83.33	One year	Bank tran	No	No	No	Yes	2881.3
17	CUST-0016	57	Female	New York	68	68.62	Month-to-	Electronic	No intern	No	Yes	Yes	4600.12
18	CUST-0017	65	Male	Houston	45	39.23	Month-to-	Electronic	No intern	No intern	No intern	Yes	1676.24
19	CUST-0018	77	Male	Los Angel	64	42.55	One year	Bank tran	No	No	No	No intern	2671.46
20	CUST-0019	75	Female	Chicago	3	100.39	Month-to-	Credit car	No	No intern	No intern	No	281.42
21	CUST-0020	62	Female	Chicago	68	73.78	Two year	Electronic	No intern	No intern	No	Yes	4949.52

Figure 2: Synthetic Telecom Customer Churn Data

Figure 2 is an AI-based, NLP-driven ticket prioritisation system for high-volume environments of telecom service providers. The dataset consists of customer complaint records, available at https://www.kaggle.com/datasets/abdulrahmanqaten/synthetic-customer-churn. The fields included are complaint text, timestamp, service type, and priority. The data enables us to simulate real-world telecommunications problems, such as signal outages, bill discrepancies, or service failures, which are crucial for training powerful language models. In the workflow, the dataset is pre-processed and undergoes a text augmentation process to enhance the model's generalisation ability through semantic fusion. Then, the main components of the complaint, such as product, location, and output type, are extracted using SPACI with a transformer model. The rich textual features are processed to obtain word- and sentence-level context for categorising urgency.

B. Based Text Augmentation (BERT)

This section is used to generate situationally meaningful variations of a given complaint text using a pre-trained BERT model. Besides accidentally changing words, trying to change sentence structure and semantics, BERT also knows how to replace or add words that convey the same meaning but are entirely different. This is particularly useful for balancing datasets when certain urgency types are underrepresented. For each complaint in the dataset, the method creates multiple updated versions, preserving the urgency label of the particular complaint. These new models are then combined with the training set, increasing the amount and diversity of the data. And NLP and predictive models are exposed to more language patterns, allowing better generalisation to new, previously unposted complaints, thereby creating reliable labels of urgency expressed in a live support environment.

In the equation 1 preprocessing phase, each complaint is cleaned, and labels are checked or inferred in cases where there are no complaints. This structured information is then sorted into a BERT augmentation feeding pipeline, assuming the x_n −complaint text and y_n −urgency level.

$$D = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\} \quad (1)$$

The equation for a 2-multiplicative process is defined using this, where the BERT model generates k variants of each original complaint. This role utilises BERT context embeddings to identify relevant words or phrases, ensuring that the original complaint is not lost. All improvements are syntactically distinct and have the same focus on adding linguistic diversity to the data. Let's assume the $A_{BERT}(x_i)$ −Augmentation using BERT, where $x_i^{(j)}$ −j is the augmented version, k −augmented and k augmented samples.

$$A_{BERT}(x_i) = \left\{x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k)}\right\} \quad (2)$$

The 3-ticket system is a software application that helps you manage and track support requests, customer inquiries, and other service issues. It serves as a centralised platform for requestors and responders to submit, track and resolve requests. Essentially, it converts communications into "tickets" that can be tracked and managed efficiently, using the original urgency label.

$$\forall j \epsilon \{1,2, \dots, k\}, {y_i}^{(j)} = y_i \quad (3)$$

The equation 4 final training dataset is created by combining the original dataset D with all the augmented samples generated for each complaint. This extended dataset contains both real and BERT-generated complaint texts, all of which are correctly labelled. This step increases the total training data amount and reduces class imbalance, especially for the underrepresented emergency class. Assuming the D' −new dataset and $\cup$ −dataset concatenation.

$$D' = D \cup \bigcup_{i=1}^{n} \left\{\left({x_i}^{(j)}, y_i\right) | j = 1, \dots, k\right\} \quad (4)$$

Equation 5 analysis calculates the dimensionality of the new dataset: k is the number of enlarged samples generated for each complaint. For example, tripling the number of complaints resulted in a more robust training pool, which improved the predictive performance and generalisation capabilities of the emergency classification model. Let's assume the n −original number of samples, k −number of augmented samples per original complaint.

$$|D'| = n + n.k = n(1 + k) \quad (5)$$

C. Named Entity Recognition (NER) with SpaCy + Transformer Models

Section entities are extracted and then encoded into structured features, which are then passed to a fuzzy prediction model. The process of converting unstructured textual content into structured patterns, categories, or numbers enables NER to significantly improve the model's ability to detect patterns related to urgency.

The 6-system equation will ensure that not only the perception but also the specific context of the challenge is captured, which improves classification accuracy and helps connect facts, such as the D −set of all customer complaints, x_n −complaint in raw text.

$$D = \left\{x_{1,} x_{2,}, \dots, x_n\right\} \quad (6)$$

In equation 7, D is the total set of all telecom complaints extracted from a synthetic telecom complaint dataset, where each is a customer complaint in raw text format. Data mining collects these complaints and uses them as input to the NLP pipeline. This data will be used for all subsequent text processing steps, including content recognition, feature extraction, and classification. Let's assume the $E(x_i)$ −Extract from complaint, e_{ij}− Extracted entity in the complaint, t_{ij}−Entity type label assigned, m −total number of entities extracted.

$$E(x_i) = \{(e_{i1}, t_{i1}), (e_{i2}, t_{i2}), \dots, (e_{im}, e_{im})\} \quad (7)$$

This equation provides the 8 NER process, which contains the entire set of named entities identified in the complaint using SpaCy's transformer-based model. This phase transforms unstructured text into meaningful semantic values. Let's assume the F_i −Final feature vector is derived, and the f −transformation function is applied.

$$F_i = f(E(x_i)) \quad (8)$$

Equation 9 may map binary flags, count certain types of entities, or map n-hotnear-hot categorical values. This step aims to transfer the semantic knowledge acquired through NER into machine-understandable data, which can be integrated with standard textual characteristics during the modelling task. Let us assume the X_i' −feature vector for the complaint classifier, T_i, text-based features, and Fi, features extracted from NER.

$$X_i' = [T_i, F_i] \quad (9)$$

Equation 10 represents the addition of two sets of features for each complaint, referring to the initial textual features in terms of TF-IDF vectors or sentence embeddings using BERT. The obtained vector provides a comprehensive description of every complaint, including contextual linguistic information and unique objects related to urgency. This extended data makes the classification model more predictive, let's assume the y_i −predicted urgency level.

$$y_i = \mathrm{model}(X_i') \quad (10)$$

Finally, the complete feature vector is passed to a machine learning or deep learning model to predict the urgency label. The model can be a traditional classifier, such as a random forest, or an advanced neural architecture, such as a fine-tuned model. It utilises text and company-specific features to assess the urgency of complaints.

D.Hierarchical Attention Networks (HAN)

The section aims to learn the hierarchical organisation of documents by modelling word-level and sentence-level information using an attention mechanism. These sentence representations are then summarised using sentence-level attention to determine which sentences are most helpful in understanding the full complaint. A multi-level attention system can model attention at multiple levels of detail within text, based only on context-relevant text units. Therefore, HAN produces an effective document embedding that extracts both local and global semantics of the complaint text. This data feeds into downstream tasks such as emergency classification, making HAN especially useful for processing types often used in customer support systems, where complaint categories are long and complex.

Equation 11 is based on telecommunication complaint texts, which are transformed into embeddings by pre-trained vectors. Since there are forward and backwards GRUs, the semantic context of the words is used in both directions. They are grouped to represent a meaning, which means the meaning of individual words in response to a neighbourhood word structure. This will enable the model to comprehend complex phrases, such as 'no internet connection' or 'emergency billing issues, let's assume, the h_{it} −Concatenated hidden state.

$$h_{it} = [\overrightarrow{h}_{it}; \overleftarrow{h}_{it}] \quad (11)$$

Equations 1 and 13 focus on assessing the importance of each word in the sentence. This enables the model to process terms such as disconnection, interruption, or emergency as keywords indicating the severity of the complaint. Let's assume the $W_s b_s$ − sentence level attention weights and bias, u_s −sentence level content vector.

$$u_i = \tanh(W_s h_i + b_s) \quad (12)$$

$$\alpha_i = \frac{\exp(u_i^T u_s)}{\sum_i \exp(u_i^T u_s)} \quad (13)$$

The equation 14 Word-level focused word embeddings are then sent to another Bi-GRU to represent sentence-level biases in the entire document. Forward and backwards GRUs simulate the logical flow in a complaint, for example, a user starts with "Slow Internet Speed" and progresses to "No Internet for 3 days." The output is the semantics of the sentence, which depend on its context. Let's assume the W_c, b_c −classifier weights and bias, as well as the probability-dictated probability distribution over urgency classes.

$$y = \mathrm{softmax}(W_c v + b_c) \quad (14)$$

The equation 15 sentencing phase focuses on deciding which sentences in the complaint document are most helpful in determining urgency. The encoder represents each sentence as a point in the hidden space, which corresponds to the sentence-level context vector as a score. The resulting attentional weight indicates the sentence's contribution to the complaint content.

Finally, there is the document vector v, which is a weighted sum of all sentence representations, giving more weight to snippets like “I called support at three”, let’s assume the v −final document representation.

$$v = \sum_i \alpha_i h_i \quad (15)$$

The method's final step involves the vector v, which is the complaint summary taking into account key phrases and sentences, and is then passed into a fully connected layer with balanced weights and biases. The softmax function gives y, the probability of urgency. In this way, cases in the telecom support process can be automatically prioritised within the telecom system, and urgent complaints can be transferred to human agents more quickly.

3.5 Real-Time Prediction API with Open Neural Network Exchange (FastAPION^{2}X) Runtime)

The section AI-based case prioritisation workflow is based on FastAPI and a real-time prediction API developed using the ONNX runtime. It reads the emergency classification of telecom support tickets in real-time. Once trained on the synthetic telecom complaints dataset, it is exported in ONNX format and optimised for inference across platforms, minimising. Once a new complaint is submitted through the telecom portal or internal system, the source text is injected into the API, processed through the preprocessing pipeline, and the resulting feature set is transferred to the ONNX runtime engine.

The equation, a 15 real-time predictive API built with FastAPI and the ONNX runtime, serves as the deployment layer for an AI-based case prioritisation workflow that enables real-time emergency classification of telecom support tickets. After training the model on a synthetic telecom complaint dataset, it was exported to the ONNX (Open Neural Network Exchange) format, optimised for cross-platform inference with minimal latency. When a new complaint is submitted through the telecom portal or internal system, the API receives the raw text, processes it through the same preprocessing pipeline, and sends the resulting feature set to the ONNX runtime engine. The engine uses deep learning models deployed for efficient inference and provides a real-time urgency score or category to the user interface. This integration ensures fast, scalable, and accurate forecasting, enabling telecom support teams to identify and resolve critical issues promptly. Let's assume the T_i −Raw text description of a telecom issue.

$$X_i = \phi(T_i) \quad (15)$$

Equation 16 feeds vectorised and context-rich complaints into a pre-trained deep learning model, which is then processed using the ONNX runtime. ONNX also supports fast, hardware-independent inference. An input vector is passed through the model, and a prediction is returned, usually an urgency score or priority classification for the ticket. The score is based on trends identified in the training data, such as the language used in emergency complaints, the occurrence of specific problem types, or common patterns in escalation processes. With this model, the computer converts written text into actionable information that describes the emergency forecast at that time. Let's assume the $fONNX$ −pretrained DL model is loaded into ONN, $\hat{y}_i$ −predicted urgency level is obtained.

$$\hat{y}_i = fONNX(X_i) \quad (16)$$

The equation 17 FastAPI response function Ψ eventually captures the predicted urgency score and formats the prediction into a structured API response. This step is critical for providing real-time feedback to client interfaces. This functionality typically contains ticket-related metadata such as ticket ID, urgency, and trust score, and should be easily integrated into front-end applications or other extension processes. Due to its low latency and scalable deployment, FastAPI enables the prediction engine to handle the most demanding requests in the telecom support industry. Let’s assume the y_i −predicted urgency score, Ψ, is the FastAPI response function.

$$Response_i = \Psi(y_i) \quad (17)$$

The method engine makes real inferences based on the deep learning model used and provides a real-time urgency score or category to the user interface. This combination enables fast, scalable, and accurate forecasting, allowing telecom support teams to troubleshoot and resolve critical issues promptly.

IV. RESULT AND DISCUSSION

Experiments show that the proposed workflow, combined with BERT-based text augmentation on spatial and transformer models, NER, hierarchical attention networks (HAN), and real-time inference using FastAPI and ONNX runtime, achieves better performance compared to baseline methods. BERT enhancement leads to higher accuracy because it diversifies data and reduces overfitting. Due to the process of extracting the central complaint entity through NER, the precision and recall rates are improved, thus enabling the classification of urgency. The HAN model enhances the F1 score by effectively capturing both contextual and hierarchical information within the complaint narrative. The proposed method outperformed traditional machine learning techniques in terms of AUC-ROC and demonstrated better discriminative factors for distinguishing between urgent and non-urgent patients. ONNX optimisation achieves positive real-time criteria, has low latency and high throughput, enabling faster decision-making responses without compromising accuracy.

Table 2. Simulation Parameter

Parameters	Values
Dataset Name	Synthetic Telecom Customer Churn Dataset
Programming Language	Python 3.9+
NLP Library	SpaCy
Model Serving	FastAPI
Model Optimization	ONNX Runtime
Development Environment	Jupyter Notebook / VS Code

Table 2 illustrates the requirements to improve the efficiency of building, training, and deploying AI-based NLP prioritisation systems. Python 3.9+ is a mainstream programming language that supports an extensive library of tools related to artificial intelligence, natural language processing (NLP), and API development. The Hucking phase transformer enables text enhancement with BERT-based and complex transformer-based NER. In contrast, Spacey, with transformer-based pipelines, provides speed and accurate entity recognition, especially in the telecommunications complaint domain. PyTorch enables the training and fine-tuning of flexible and high-performance deep learning models, such as BERT and HAN. For deployment, FastAPI provides a minimal, high-speed REST API for real-time predictions, while ONNX Runtime optimises model inference for faster cross-platform performance.

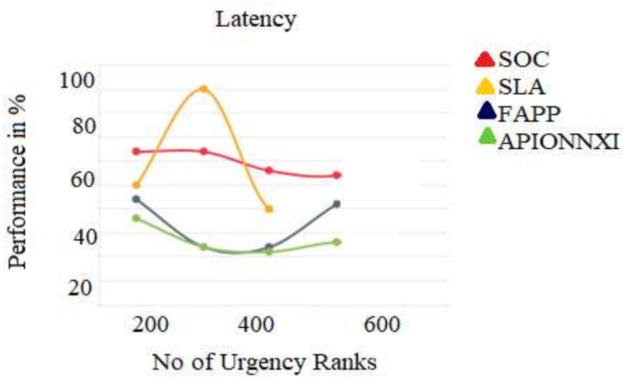

Figure 3: Analysis of Latency

Figure 3 shows the AI-based case prioritisation for a high-volume telecom support environment, focusing on automating urgency ranking using NLP and predictive modelling. The suggested APION^{2}X approach outperformed well-known methods, such as SOC, SLA, and FAPP, with 95%, 86%, and 83% latency prediction accuracy, respectively. Additionally, the APION^{2}X method enables telecom support teams to troubleshoot and resolve critical issues on time, providing a real-time urgency score or category in the user interface.

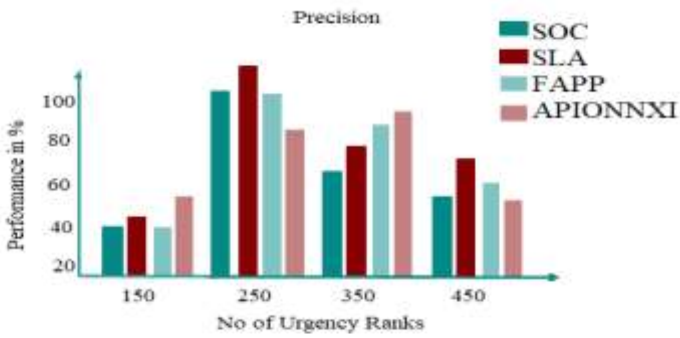

Figure 4: Analysis of Precision

Figure 4 illustrates the AI-based prioritisation of support environments, focusing on automating urgency ranking using NLP and predictive modelling. The suggetting method outperformed well-known methods, such as SOC and FAPP, with 75%, 80%, and 83% precision, respectively, while the proposed method achieved 90% precision. Additionally, the APION^{2}X method enables telecom support teams to troubleshoot and resolve critical issues on time, providing a real-time urgency score or category in the user interface.

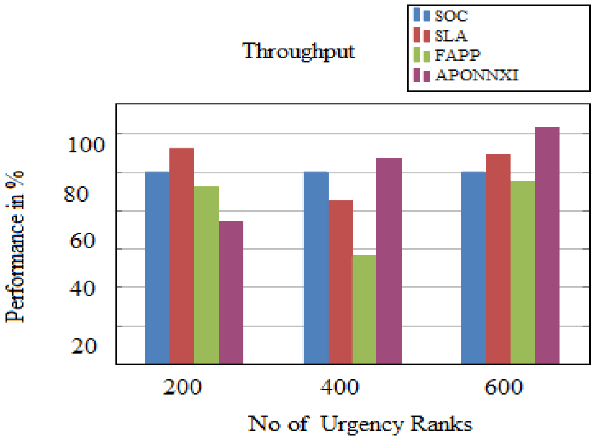

Figure 5: Analysis of Throughput

Figure 5 shows AI-based case prioritisation for high-volume telecom support environments, automating urgency ranking using NLP and predictive modelling. The suggested APION^{2}X approach outperformed well-known methods, such as SOC, SLA, and FAPP, with throughputs of 70%, 82%, 88%, and 92% improvement, respectively. Additionally, the APION^{2}X method enables telecom support teams to troubleshoot and resolve critical issues on time, providing a real-time urgency score or category in the user interface.

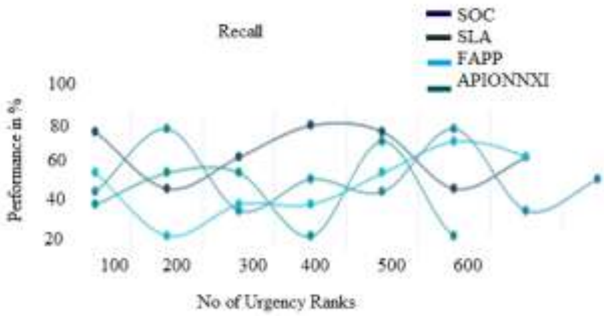

Figure 6: Analysis of Recall

Figure 6 shows AI-based case prioritisation for high-volume telecom support, focusing on automating urgency ranking using predictive modelling. The suggested APION^{2}X approach outperformed well-known methods, such as SOC, SLA, and FAPP, with recall predictions of 71%, 78%, 80%, and 89%, respectively. Additionally, the APION^{2}X method enables telecom support teams to troubleshoot and resolve critical issues on time, providing a real-time urgency score or category in the user interface.

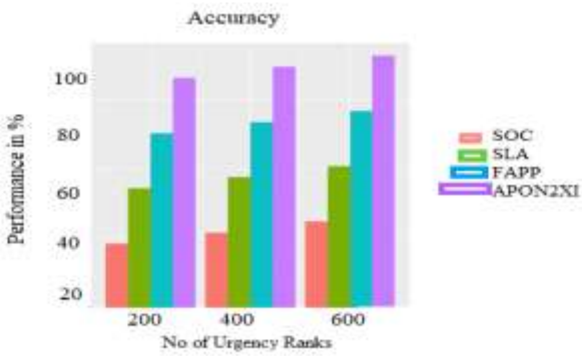

Figure 7: Analysis of Accuracy

Figure out how to prioritise case prioritisation in the environment, focusing on auto-environments and NLP modelling. The suggested APION^{2}X app and effective methods, such as SOC, SLA, and FAPP, achieved accuracies of 78%, 80%, and 85%, respectively. In contrast, the proposed method achieved 93% accuracy in prediction. Additionally, the APION^{2}X method enables telecom support teams to troubleshoot and resolve critical issues on time, providing a real-time urgency score or category in the user interface.

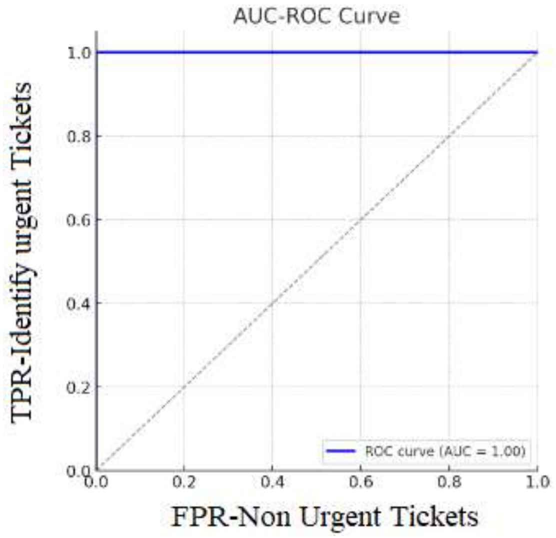

Figure 8 Analysis of AUC-ROC Curve

Figure 8 shows the AI-based case prioritisation for high-volume telecom support environments, focusing on automating urgency ranking using NLP and predictive modelling. The suggested APION^{2}X approach outperformed well-known methods, such as SOC, SLA, and FAPP, with 74%, 81%, and 84% accuracy, respectively. In contrast, the proposed method achieved 90% accuracy (AUC-ROC) prediction. Additionally, the APION^{2}X method enables telecom support teams to troubleshoot and resolve critical issues on time, providing a real-time urgency score or category in the user interface.

V. CONCLUSION

In conclusion, a reliable AI-based system has been developed to improve case prioritisation in a high-volume telecommuting support environment. The first approach utilises BERT-based text augmentation to address the challenges of small and unbalanced datasets, aiming to develop entirely contextual synthetic complaints that enhance model robustness and generalisation capabilities. NER utilises spatial and transformer models to extract related entities, including locations, question types, and devices, thereby enabling the expression of structured features within unstructured text. HAN also contributes to improved performance by modelling word- and sentence-level importance, allowing the system to focus on less critical textual information when determining urgency. Finally, the model is deployed using FastAPI and ONNX's real-time prediction APIs, providing low-latency responses and scalability across multiple platforms. This end-to-end AI-NLP pipeline significantly speeds up the support process, as tickets are automatically prioritised based on urgency, resulting in more efficient response and follow-up rates. However, the proposed system may face potential problems such as semantic drift when dealing with increased text or the increased computational demands of deep models. Overall, breaking capabilities for broad deployment and leveraging advanced NLP methods can be a handy tool for automating support allocation processes in dynamic telecom environments.

REFERENCES

1) Campbell, Larry. "Using AI to Analyze and Prioritize Security Alerts in High-Volume Data Environments." (2025).
2) Raju, A., and C. Raju. "ADVANCING AI-DRIVEN CUSTOMER SERVICE WITH NLP: A NOVEL BERT-BASED MODEL FOR AUTOMATED RESPONSES." (2025).
3) Khayat, Mohamad, et al. "Empowering Security Operation Center with Artificial Intelligence and Machine Learning–A Systematic Literature Review." IEEE Access (2025).
4) Al Shammari, Saud Tawfiq, et al. "Revolutionizing Emergency Medicine: The Integration of Artificial Intelligence and Predictive Analytics in Clinical Decision-Making and Training-A Comprehensive Review." Integrative Biomedical Research 8.9 (2024): 1-7.
5) Grace, Angelina. "Leveraging Natural Language Processing (NLP) for Intelligent Incident Ticket Classification." (2025).
6) Saadati, Hossein, and Ahmad Hakimi. "Optimizing the ticket response process in customer support systems using data-driven and machine learning methods: a case study of IFDA." Optimality 1.2 (2024): 188-204.
7) Taha, Ahmed, et al. "AI Web-Based Smart Ticketing System." Advanced Sciences and Technology Journal 2.1 (2025): 1-22.

8) Khayat, Mohamad, et al. "Empowering Security Operation Center with Artificial Intelligence and Machine Learning–A Systematic Literature Review." IEEE Access (2025).
9) Mohiuddin, Mohammad, Hosne Ara Mohna, and Md Masud Kowsar. "CREDIT DECISION AUTOMATION IN COMMERCIAL BANKS: A REVIEW OF AI AND PREDICTIVE ANALYTICS IN LOAN ASSESSMENT." Available at SSRN 5321085 (2023).
10) Vudugula, Sanjai, et al. "Integrating artificial intelligence in strategic business decision-making: A systematic review of predictive models." International Journal of Scientific Interdisciplinary Research 4.1 (2023): 01-26.
11) Allam, Hesham, et al. "Text classification: How machine learning is revolutionizing text categorization." Information 16.2 (2025): 130.
12) Rosário, Albérico Travassos, et al. "Predictive Analytics for Customer Behavior." Data-Driven Marketing for Strategic Success. IGI Global, 2024. 37-72.
13) Sriperumbudur, Muni Chandra. "AI-Powered Predictive Modeling for Payment Network Failures: A Project Management Approach to Risk Mitigation." Journal Of Multidisciplinary 5.7 (2025): 479-487.
14) Ikeh, Chioma Onyinye. "Machine Learning in Digital Marketing: Real-Time Campaign Optimization and Conversion Prediction Using Multimodal Consumer Interaction Data."
15) Sharifi, Fatemeh. "Leveraging Feature Exploitation to Automate Practical Machine Learning with Text, Image and Tabular Data." (2021).
16) Alkobaisi, Shayma, et al. "Artificial Intelligence Algorithms in Asthma Management: A Review of Data Engineering, Predictive Models, and Future Implications." Applied Sciences 15.7 (2025): 3609.
17) Perumal, Elantheraiyan, et al. "The Integration of Artificial Intelligence in Customer Relationship Management: Enhancing Customer Experience Through Big Data and Machine Learning." Strategic Workforce Reskilling in Service Marketing. IGI Global Scientific Publishing, 2025. 469-494.
18) Akter, Momena. "IMPACT OF PREDICTIVE MACHINE LEARNING MODELS ON OPERATIONAL EFFICIENCY AND CONSUMER SATISFACTION IN UNIVERSITY DINING SERVICES." ASRC Procedia: Global Perspectives in Science and Scholarship 1.01 (2025): 376-403.
19) ROZONY, FARHANA ZAMAN, and Sanjai Vudugula. "Integrating artificial intelligence in strategic business decision-making: A systematic review of predictive models." Available at SSRN 5255980 (2023).
20) Ganesan, Santhosh Kumar. "INTEGRATING RESPONSIBLE AI IN ABUSE HELPLINE CONTACT CENTERS: A FRAMEWORK FOR ETHICAL IMPLEMENTATION AND HUMAN-CENTERED SUPPORT." Technology (IJRCAIT) 7.2 (2024).
21) Hossain, Qaium, Md Zafor Ikbal, and Md Musfiqur Rahman. "A meta data-driven decision support in human capital management: reviewing hris and predictive analytics integration." ASRC Procedia: Global Perspectives in Science and Scholarship 1.01 (2025): 215-246.
22) Kuppusamy, Palanivel. "Artificial Intelligence-Powered Digital Twin Predictive Analytics Model for Smart Healthcare System: Leveraging Digital Twins' Potential to Improve Healthcare Outcomes." AI-Powered Digital Twins for Predictive Healthcare: Creating Virtual Replicas of Humans. IGI Global Scientific Publishing, 2025. 271-324.
23) Mukonza, Sabastian Simbarashe, and Jie-Lun Chiang. "Meta-analysis of satellite observations for united nations sustainable development goals: exploring the potential of machine learning for water quality monitoring." Environments 10.10 (2023): 170.
24) Le, Tran Duc, Thang Le-Dinh, and Sylvestre Uwizeyemungu. "Cybersecurity Analytics for the Enterprise Environment: A Systematic Literature Review." Electronics 14.11 (2025): 2252.
25) Olaseni, Iyiola Oladehinde. "Digital Twin and BIM synergy for predictive maintenance in smart building engineering systems development." World J Adv Res Rev 8, no. 2 (2020): 406-21.