The Eurasia Proceedings of
Educational & Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational and Social Sciences (EPESS), 2025

Volume 46, Pages 125-136

IConETE 2025: International Conference on Education in Technology and Engineering

Classification of User Complaints and Requests in Customer Service Using Bert Fine-Tuned with Lora

Simge Senyuz
Aktif Investment Bank Inc.

Erkut Baloglu
Aktif Investment Bank Inc.

Eren Caglar
Aktif Investment Bank Inc.

Ismail Gocmez
Aktif Investment Bank Inc.

Abstract: The categorization of user complaints and requests is an essential task for large companies, as user feedback is crucial for customer satisfaction and business development. However, manual categorization of such data is highly labor intensive and time consuming. To address this, we automated the classification using classical ML methods and BERT on over 18,000 user complaints and requests spanning 11 classes. We utilized LoRA to make the BERT fine-tuning more computationally efficient by reducing trainable parameters while preserving performance. Given the dataset imbalance, we augmented the minority classes with paraphrasing via gemma-7b. The fine-tuned BERT achieved a 5-10% performance improvement over traditional machine learning approaches, including logistic regression, SVM, and XGBoost, and showed robust results exceeding 80% in all key metrics (82% accuracy, 81% precision, 82% recall, and 81% F1-score) on the test set. This work highlights the potential to reduce manual labor costs while ensuring high accuracy in real-world customer service applications.

Keywords: Machine Learning, Fine-tuning, BERT, LoRA, Customer support

Introduction

With millions of user interactions daily, large firms face a critical challenge in efficiently categorizing user complaints and requests. However, classifying such data by hand takes a lot of time and effort. Automated classification systems are applied to prioritize urgent complaints, route requests to appropriate agents, identify complaint severity, and extract root causes (Vairetti, 2024; Poczeta, 2023; Song, 2024). These systems support business process improvements, risk assessment, and tailored managerial responses, leading to increased user satisfaction and reduced churn.

Recent research leverages machine learning (ML), deep learning (DL), and natural language processing (NLP) techniques to categorize, prioritize, and analyze complaints across industries. Modern approaches use transformer-based models (e.g., BERT) (Vairetti, 2024; Xu, 2025) and deep learning architectures (e.g., BiLSTM-CRF, Doc2Vec) (Vairetti, 2024). For high-accuracy complaint categorization and prioritization, often outperforming traditional methods like SVM, TF-IDF, and bag-of-words. Domain-specific training and hierarchical label structures further

enhance performance (Zangari, 2023; Kelm, 2024). There are different approaches for hierarchical classification: 1) per-level classifiers, where each level of the hierarchy is classified individually 2) per-parent classifiers where hierarchy is included in the classification process. In the per-parent classification method, a separate classifier is trained for each main class (Zangari, 2023).

To our knowledge, all of the aforementioned models were completely trained or optimized. However, when the size of the training dataset increases, training a BERT model or a comparable architecture becomes ineffective, time-consuming, and computationally costly. Large pre-trained models can now be effectively adapted to domain-specific tasks while using less computing power and training time thanks to recent developments in parameter-efficient fine-tuning, especially Low-Rank Adaptation (LoRA) (Hu, 2022).

Here, we analyzed complaints and requests of Aktif Bank and N Kolay. First, a subset of the dataset containing 18,137 samples and 11 imbalanced classes was classified using logistic regression, SVM, XGBoost, and BERT fine-tuned with LoRA. In order to address the class imbalance problem, we paraphrased the minority classes using gemma-7b (Gemma Team, 2024). The fine-tuned BERT achieved a 5-10% performance improvement over traditional ML approaches and showed robust results exceeding 80% in all key metrics (82% accuracy, 81% precision, 82% recall, and 81% F1-score) on the test set. Second, as the best-performing method is BERT, we applied BERT classifiers on the full dataset, which has 48,785 samples spanning 11 main classes (hereby referred to as Level 1 classes) and 165 sub-classes (hereby referred to as Level 2 classes). For Level 2 classes, a different classifier head was trained for each of the corresponding Level 1 classes. In this dataset, Level 1 performance is around 80% across all metrics, whereas Level 2 performance is around 70%.

Literature Review

Application of Text Classification in Customer Support

There is extensive research on the applications of text classification in customer support. The research focuses on various areas such as effective categorization of users' complaints, understanding the urgency or severity of a complaint, and ticket automation.

Vairetti et al. (2024) used multicriteria decision-making (MCDM) and DL to automatically prioritize users' complaints. By integrating transformer-based models (especially BERT versions) for text classification with MCDM techniques to generate training labels, the authors tackle the problem of manually classifying and ranking high volumes of complaints. In order to create binary labels that differentiate between urgent and non-urgent complaints, they use data from a Chilean occupational safety organization that has received over 44,000 complaints and inquiries. They do this by using a weighted-sum MCDM approach that takes into account eight factors, such as the nature of the claim, the service provided, and the priority level. The BERT model trained on Spanish text achieved the greatest performance with 92.1% accuracy and 0.9785 AUC. For business analytics applications, this framework offers a unique fusion of artificial intelligence (deep learning) and operational research (MCDM).

Advanced NLP strategies have been investigated in recent automated ticket classification research to enhance support ticket categorization (Zangari, 2023). examined the use of transformer-based language models for hierarchical ticket classification tasks and carried out a thorough analysis of ticket automation techniques. The authors showed that document embedding strategies have a considerable impact on classification performance using datasets of financial customer complaints and Linux bug reports. Their ideal strategy outperformed traditional methods by more than 28% in terms of F1-score compared to the standard methods. By merging classifiers trained on several levels of the label hierarchy, the study presented multi-level classification frameworks, ML-BERT and SupportedBERT, which make use of hierarchical label structures. With an F1-score improvement of up to 5.7%, their strategies beat conventional baseline techniques like SVMs and recurrent neural networks. For efficient ticket automation in actual customer support systems, the authors came to the conclusion that carefully choosing document representation schemes and incorporating hierarchical information into categorization designs are essential components.

Xu et al. examines the effects of various user complaints in online hotel reviews on overall satisfaction by combining DL and econometric modeling (Xu, 2025). From more than 400,000 evaluations of Beijing hotels on Ctrip.com, the authors automatically identify seven complaint areas (service, facility, cleanliness, price, location, dining, and noise)

using a hybrid BERT-BiLSTM-CRF model. They achieve an F1 score of 0.82 and a recall of 0.85. According to their econometric analysis, the impact of various complaint categories on user satisfaction varies; the most detrimental effects are seen in complaints about cleanliness and service, followed by those about facilities, costs, and noise. It's interesting to note that satisfaction and dining complaints are positively correlated, which the authors attribute to a decline in users' expectations for value-added services. By applying dominance analysis, they show that while the top three complaint categories (cleanliness, facility, and service) account for 86.2% of the variance in satisfaction, service concerns contribute 45.9%. The results hold up well both before and during the COVID-19 epidemic; however, reviews conducted during that time period revealed a greater focus on issues related to price, cleanliness, and service. This gives hotel managers useful information to help them prioritize service quality enhancements. When taken as a whole, these works show how the field has developed along three axes: methodological integration with allied fields, architectural innovation, and practical deployment that tackles real-world limitations like complex taxonomies and label scarcity.

Overview of Augmentation Methods Used in Imbalanced Text Classification

Imbalanced data is a major challenge in text classification, including customer support, where certain categories such as rare complaint types are underrepresented. Recent studies have developed and evaluated a range of methods (e.g., data augmentation, ensemble learning, deep learning, and hybrid sampling) to address this issue and improve classification performance. There are some effective methods that tackle this problem, like data augmentation methods using SMOTE, adaptive synthetic sampling, GAN, BERT, or LLMs such as GPT (Khan, 2024). These methods are often combined with ensemble methods such as bagging and boosting. In this comparative analysis, Khan et al. evaluated the performance of several ML methods, including XGBoost, LightGBM, and AdaBoost, combined with data augmentation methods such as SMOTE and GAN, on an imbalanced dataset. Their findings show that SMOTE-LightGBM and ROS-LightGBM outperform other combinations. The authors also emphasize that using complex methods such as GAN is both complicated in terms of hyperparameter tuning and computationally expensive.

In order to overcome class imbalance in text classification problems, Taskiran et al. provides a thorough benchmarking of SMOTE and thirty of its variations (Taskiran, 2025). After vectorizing the text using the MiniLMv2 transformer model, they applied these oversampling techniques to two benchmark datasets (TREC and Emotions). While techniques such as Polynom Fit, SMOTE, and DE Oversampling enhanced performance on the TREC dataset, many techniques demonstrated limited generalization ability on the more difficult Emotions dataset, especially with Decision Trees and Random Forest displaying overfitting. In addition to highlighting the importance of carefully evaluating dataset characteristics, class distribution patterns, and the intrinsic sensitivities of various classifiers to synthetic sample generation, the study uses Friedman testing to validate statistical significance.

In addition to the classical SMOTE and similar approaches, using LLMs as paraphrasers has been shown to be a reliable method to tackle the imbalanced class problem (Abaskohi, 2023; Yadav, 2024; Zhang, 2024). Paraphrasing minority class samples using LLMs (e.g., GPT-3, T5, LLaMA, ChatGPT) is widely used to generate synthetic data, effectively balancing class distributions and improving classifier performance.

PAG-LLM (Paraphrase and Aggregate with Large Language Models), a novel approach for reducing errors in large-scale intent classification tasks. In this work, the authors integrated the paraphrasing into the classification process itself. They applied a three-stage process where an LLM generates multiple paraphrases of input queries, performs classification on both the original and paraphrased versions, and aggregates results based on confidence scores. The method is particularly effective for low-confidence predictions, where uncertainty is highest, and can be selectively applied to only 32% of test inputs while still achieving substantial improvements. Evaluated on two large intent classification datasets (CLINC with 150 classes and Banking with 77 classes), PAG-LLM achieves 22.7% and 15.1% error reduction, respectively, with significant improvements in both misclassification errors and hallucinated label generation.

According to the literature, different augmentation techniques for imbalanced text categorization have different trade-offs. Taskiran et al. (2025) show that traditional SMOTE-based methods provide strong performance and computational efficiency, especially when paired with ensemble methods like LightGBM (Khan et al., 2024). However, they are susceptible to overfitting and have limited generalization on complex datasets.

Despite their theoretical complexity, GAN-based techniques require a lot of processing power and hyperparameter optimization, and they do not consistently beat simpler SMOTE equivalents (Khan, 2024). By contrast, a paradigm change toward semantically richer augmentation is represented by LLM-based paraphrasing employing models such as GPT-3 and T5 (Abaskohi, 2023; Yadav, 2024; Zhang, 2024). LLM-based approaches provide better semantic comprehension for complicated, multi-class problems, even while older methods are still effective in situations with limited resources.

Methodology

The workflow of hierarchical classification is given in Figure 1. The input texts were classified as Level 1 and Level 2 classes. A BERT model was trained on Level 1 classes; then, a Level 2 classifier was trained for each Level 1 class separately.

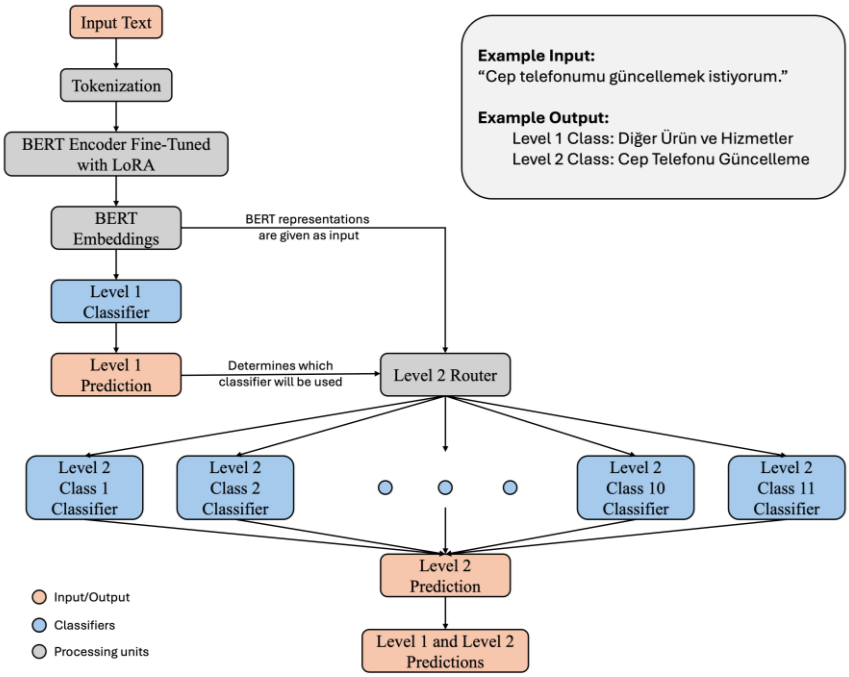

Figure 1. Hierarchical classification workflow

Data Collection and Augmentation

Users' complaints and requests were collected over a 3-month period from Aktif Bank and N Kolay. These complaints and requests come from different resources, such as the web forms that users filled out or were created by a customer support agent at the end of a call coming from the user. Since some of these were filled by the users themselves, there is a possibility of noise in the labelling of the dataset. After the datasets were collected, stopwords were removed from the samples, the dataset was randomized, and an 80%-20% stratified train-test split was applied. Relevant functions of scikit-learn were used to perform stratification and splitting. To mediate the imbalance problem, the minority classes were paraphrased using gemma-7b (Gemma Team, 2024). The augmented samples were included in the training but not the test set to keep the test performance as realistic as possible.

Classification with Classical ML Methods

For the classification of users' complaints and requests, we utilized logistic regression, SVM, and XGBoost. First, we used mxbai embedding through Ollama (accessible through github.com/ollama/ollama) to turn the texts to embeddings. Mxbai embeddings were chosen since they support non-English texts even though they weren't trained

on Turkish specifically and based on its strong performance on the MTEB benchmark (64.68 average across 56 datasets), particularly its high scores in classification tasks (75.64), and semantic textual similarity (85.00) (huggingface.co/mixedbread-ai/mxbai-embed-large-v1). In addition, these embeddings are easy to integrate with Ollama. After the embeddings were created, Level 1 classes were predicted using logistic regression, SVM, and XGBoost. Scikit-learn was used for the predictions implementation of these methods. For logistic regression, the lbfgs solver was used with 1000 max iterations. A multinomial loss function was used. For SVM, the LinearSVC function of scikit-learn was used, and class weights were set to balanced. For XGBoost, we employed a multiclass soft probability objective (multi:softprob) with the following hyperparameters: learning rate of 0.1, 200 boosting rounds, maximum tree depth of 6, row subsample ratio of 0.8, and column subsample ratio of 0.8. Model performance was evaluated using multi-class logarithmic loss.

Classification with BERT Fine-tuned with LoRA

A BERT model fine-tuned with Turkish texts was used as the baseline model (accessible through huggingface.co/dbmdz/bert-base-turkish-cased). The model was fine-tuned with LoRA (Hu, 2022) using the transformers module. Only the attention layers of the BERT model were used in LoRA, whereas other layers were frozen. We optimized various hyperparameters of LoRA: rank (r = 4, 8, 16, 32), alpha (lora_alpha = 8, 16, 32, 64), and dropout (lora_dropout = 0.1, 0.5). The BERT model was trained for 10 epochs with early stopping with a patience of 3 epochs. The learning rate was chosen as 2e-5, AdamW was used as the optimizer, and cross entropy loss was used as the loss function. The model was trained using transformers, torch, and peft modules of Python. For the Level 1 classification, a single classifier head was used, whereas for the Level 1 + Level 2 classification, separate classifier heads were utilized for each Level 1 class.

Evaluation

We used accuracy, precision, recall, and F1-score metrics to evaluate the performance of the models. Throughout the paper, the weighted averages of these scores were reported, meaning that the averages were normalized based on the number of samples for each class. For Level 2 metrics, averages were taken within the corresponding Level 1 class. Accuracy was calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision was calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

Recall was calculated as follows:

$$Recall = \frac{TP}{TP + FN}$$

F1-score was calculated as a combination of the precision and recall as follows:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

For the evaluation of each method, confusion matrices were created using the metrics above. These confusion metrics were row normalized in order to compare the performance of the model in different classes.

Results

Data Description

User complaints and requests were classified into two different subsets, one having 18,137 samples and the other having 48,785 samples. On average, there are 16 words in the samples of the first dataset, whereas the longest text has 434 words. These datasets have 11 Level 1 classes and 165 Level 2 classes. Datasets of user requests and complaints are inherently quite imbalanced. The most populated three classes comprise 36% (Other Products and Services), 24% (Credits), and 11% (Banking Channels), over 70% of the whole dataset (Table 1). Level 2 classes are also imbalanced. For example, the Level 1 Credits class has 37 unique Level 2 classes, whereas Invoices and Payments has only 4. The performance of every model we trained suffers as a result of these imbalances, with underrepresented classes doing worse and the most populous classes performing exceptionally well (with over 90% accuracy).

Table 1. Class distribution of Level 1 classes and number of Level 2 classes per Level 1 classes

Class	Number of Samples	Number of Level 2 Classes per Level 1 Classes
Banking Channels	5313	23
Other Products and Services	17435	20
Invoices and Payments	119	4
Service Quality	711	18
Cards	4821	21
Credits	11730	37
Money Transfer Issues	2452	8
Insurance	1366	9
Checking Account Issues	2702	8
Inheritance Procedures	1285	6
Investment Products and Services	851	11

Performance Comparison of Classical ML Method vs BERT

We classified Level 1 classes using logistic regression, SVM, XGBoost, and BERT fine-tuned with LoRA on the smaller dataset. The best-performing method was selected for the hierarchical classification. The results of logistic regression (Figure 2A), SVM (Figure 2B), and XGBoost (Figure 2C) are given. The results of the fine-tuned BERT model are given in Figure 3.

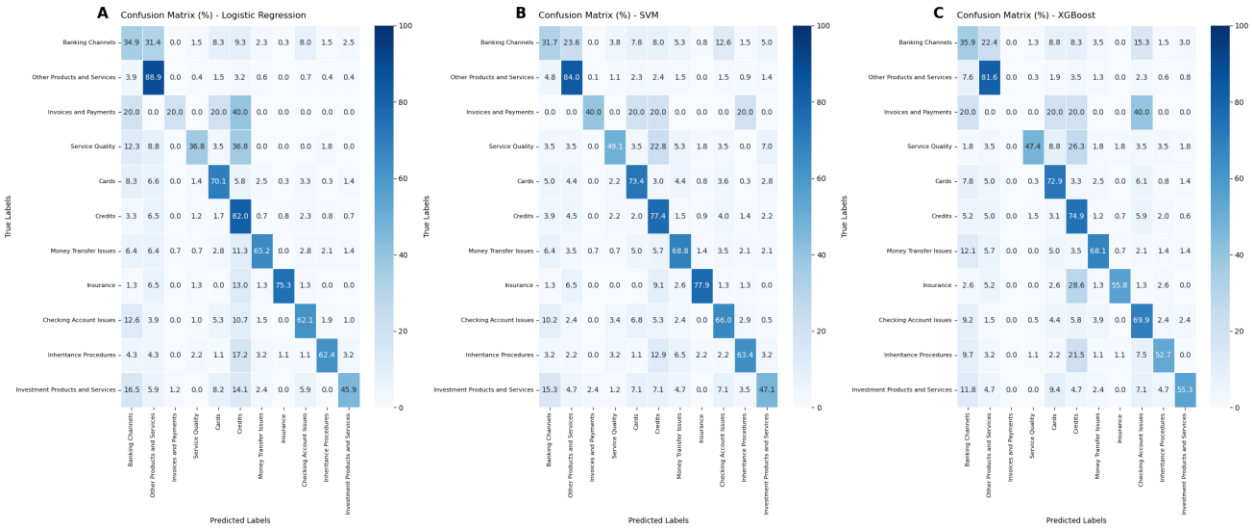

Figure 2. Confusion matrices of Logistic Regression, SVM, and XGBoost

The BERT model outperforms other models in all of the classes (Figures 2 and 3). There is an approximately 5-10% difference in true positive percentages in all classes. ML methods perform acceptably in highly populated classes such as Other Products and Services. However, BERT shows exceptional performance even in the classes with a lower number of samples, such as Inheritance Procedures. All of the models perform poorly in Invoices and Payments, which has the lowest number of samples.

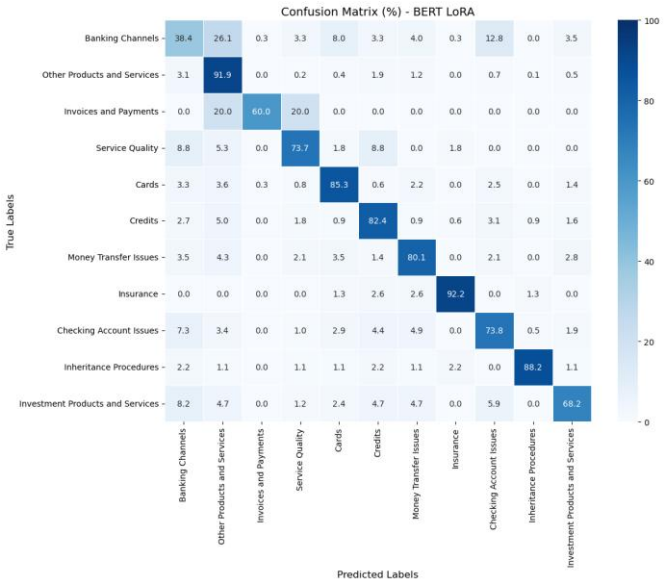

Figure 3. Confusion matrix of fine-tuned BERT model

Hierarchical Classification using BERT

As the best-performing model is the BERT model fine-tuned with LoRA, we applied this model in the hierarchical classification. The BERT model exhibits 83% accuracy for Level 1 classifications and 70% accuracy for Level 2 categories (Table 2 and Figure 4). The average accuracy of Level 2 rises to 84% if Level 1 predictions are accurate, indicating that Level 1 predictions clearly form a bottleneck for our classification problem.

Table 2. The Level 1 performance of hierarchical BERT classifier on the test set

Class	Precision	Recall	F1-Score	Number of Test Samples
Banking Channels	0.60	0.43	0.50	1063
Other Products and Services	0.87	0.93	0.90	3487
Invoices and Payments	0.59	0.67	0.63	24
Service Quality	0.85	0.46	0.60	142
Cards	0.81	0.84	0.83	964
Credits	0.86	0.88	0.87	2346
Money Transfer Issues	0.79	0.84	0.82	490
Insurance	0.91	0.93	0.92	273
Checking Account Issues	0.59	0.67	0.63	541
Inheritance Procedures	0.93	0.86	0.89	257
Investment Products and Services	0.78	0.51	0.62	170

Banking Channels, Other Products and Services, and Checking Account Issues are misclassified at a high rate on Level 1 (Figure 4). Furthermore, Service Quality and Credits are misclassified as one another. The largest misclassification is between Checking Account Issues and Banking Channels, with 41.63% of Channel Blocks cases (87 cases out of 209 samples) being mistakenly classified as Account Blocking Procedures. Similarly, Banking Channels cases are often misclassified as Other Products and Services, with substantial confusion in categories such as OTP - Soft OTP (69.35% error rate, 43 instances) and Registration-Login Issues (28.65% error rate, 100 cases) (Table 3). The linguistic similarity between these groups makes classification challenging for the models.

When the Level 1 prediction is accurate, the model shows a respectable level of accuracy inside the Other Products and Services class, properly classifying the majority of cases into their Level 2 subcategories (Table 3). A worrying trend, however, shows that 49 out of 1,475 samples, or 3.32% of Alternative Security Question Unblocking Process cases, are incorrectly classified at Level 1 as Banking Channels, more precisely as Registration-Login Issues. The

semantic overlap between security-related processes and login actions is indicated by this cross-category confusion. The Credits category also exhibits a moderate level of misunderstanding with Cards, with 17.42% of Application Inquiries being incorrectly categorized as Credit Restructuring Processes (27 instances).

Table 3. Top 10 misclassified categories

Level 1 True Label	Level 2 True Label	Level 1 Predicted Label	Level 2 True Label	Number of Incorrect Predictions	Number of Samples	Percentage of Misclassifications
Banking Channels	Registration-Login Issues	Other Products and Services	Alternative Security Question Unblocking Process	100	349	28.65
Banking Channels	Channel Blocks	Checking Account Issues	Account Blocking Procedures	87	209	41.63
Other Products and Services	SIM Card Unblocking	Other Products and Services	Alternative Security Question Unblocking Process	65	442	14.71
Other Products and Services	Corporate Password-Unblocking and Update Process	Other Products and Services	Alternative Security Question Unblocking Process	63	185	34.05
Other Products and Services	Mobile Phone Number Update	Other Products and Services	Alternative Security Question Unblocking Process	58	1082	5.36
Other Products and Services	Alternative Security Question Unblocking Process	Banking Channels	Registration-Login Issues	49	1475	3.32
Checking Account Issues	Account Blocking Procedures	Banking Channels	Channel Blocks	45	261	17.24
Banking Channels	OTP - Soft OTP	Other Products and Services	Alternative Security Question Unblocking Process	43	62	69.35
Credits	Application Inquiries	Credits	Credit Restructuring Processes	27	587	4.60
Cards	Card Delivery and Courier Routing Issues	Cards	Application Inquiries	27	155	17.42
Credits	Application Inquiries	Credits	System -Screen Problems	27	587	4.60

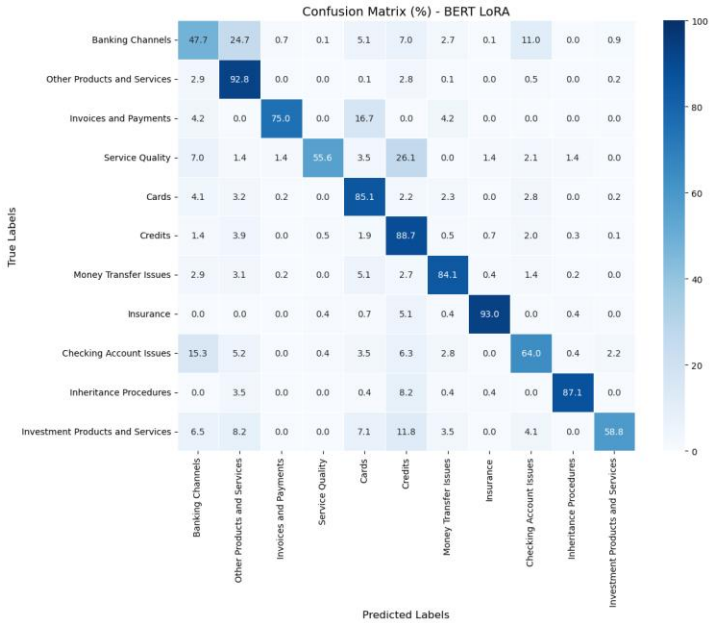

Figure 4. Confusion matrix of Level 1 classes of hierarchical classification. The values are normalized by row

Discussion

Every year, financial institutions handle millions of consumer requests and complaints, posing a significant operational issue that has an immediate effect on corporate efficiency, regulatory compliance, and consumer satisfaction (Bastani, 2019). When critical complaints are not handled right away, manual categorizing of these encounters is time-consuming, labor-intensive, and prone to inconsistencies (Yilmaz, 2016). This can result in irate consumers, higher churn rates, and possible regulatory infractions. The complexity is increased in banking settings as complaints cover a wide range of service sectors, from digital channels and security protocols to account operations and credit products, all of which need to be forwarded to specialist teams with unique specialties. In contrast to accurate automated classification, which allows for fast routing to the relevant agents, prioritizing of important issues, and data-driven business process improvements, misclassified complaints lead to numerous handoffs between departments and prolonged resolution periods. Traditional classification algorithms are unable to effectively handle the layers of complexity added by the hierarchical nature of financial services, where broad categories contain several specialized subcategories (Roumeliotis, 2025). Furthermore, classification systems must function consistently across regular requests with large volume and low frequency but possibly high impact complaints due to the inherent class imbalance in complaint data, where some concerns are overrepresented while significant but infrequent problems are underrepresented.

In this study, we analyzed the user complaints and requests that came to Aktif Bank and N Kolay. We developed and evaluated an automated hierarchical classification system for complaints and requests using BERT fine-tuned with LoRA. Our findings provide actionable insights for both technical system improvements and business process optimization in automated complaint management systems.

The dataset has an inherent imbalance (Table 1) that significantly affects model performance for all investigated architectures; the most populous categories consistently achieve higher accuracy than 90%, while underrepresented classes achieve lower accuracy (Table 2). In order to overcome this difficulty, we used gemma-7b paraphrasing approach to enhance minority classes; nevertheless, this method resulted in only slight gains (1–2% across key metrics) and a considerable increase in training time. If sufficient minority class samples can be collected, undersampling the minority classes may be used in future research.

Using a small dataset for Level 1 classification, we compared the BERT model with the conventional ML techniques (logistic regression, SVM, and XGBoost). The best-performing model was chosen for further hierarchical

classification (as described in Figure 1) on the entire dataset after the outcomes of logistic regression (Figure 2A), SVM (Figure 2B), and XGBoost (Figure 2C) were contrasted with the optimized BERT model with LoRA (Figure 3). The BERT model performed better than the classical ML methods across all classes, with robust results exceeding 80% in all key metrics (82% accuracy, 81% precision, 82% recall, and 81% F1-score) on the test set.

It also achieved roughly 5–10% higher true positive rates (Figures 2 and 3). Even though classical ML techniques produced results that were satisfactory for majority classes like Other Products and Services, the performance disparity was especially noticeable in minority classes. BERT shown remarkable performance in low-sample categories such as Inheritance Procedures, where traditional approaches had considerable difficulties. All models, however, had significant issues with Invoices and Payments, the class with the lowest number of samples, underscoring the basic problem of severe class imbalance that cannot be entirely addressed by model architectural enhancements alone.

Despite the clear performance advantages of BERT, classical ML methods offer significant computational benefits, requiring considerably fewer resources than BERT fine-tuning. By limiting the trainable parameters to just 1–1.5% of the entire network, LoRA's approach mitigates the computational cost and significantly improves the viability of BERT fine-tuning in commercial settings. However, in situations where real-time classification is necessary or where deployment resources are limited, a hybrid architecture that uses BERT for minority class predictions and ML techniques for majority class predictions may offer the best compromise between computational efficiency and classification accuracy.

Due to its better performance, the BERT model with LoRA was chosen for hierarchical classification on the entire dataset, which included 48,785 samples. Distinct classifier heads were trained for each associated Level 1 class in order to predict Level 2 labels. The findings of the hierarchical classification provide important new information about our system's mistake patterns and performance problems. The conditional accuracy of 84% for Level 2 predictions (assuming correct Level 1 classification) indicates that Level 1 misclassifications are the main limiting factor in system performance, even though the model achieves 83% accuracy at Level 1 and 70% overall accuracy for the entire two-level hierarchy.

The model performs differently in each Level 1 category, with well-populated classes typically showing good performance (Figure 4). It's interesting to note that some of the other classes with comparatively small sample sizes also perform well, such as Insurance and Inheritance Procedures. The semantically unique nature of these categories, which include specific vocabulary and concepts that are easily distinguished from other complaint types, is probably the reason for this good performance even with smaller sample sizes.

In the case of underperforming categories like OTP - Soft OTP (69.35% misclassification rate) and Invoices and Payments, the low sample counts and high confusion rates indicate that these problems might be underreported or inadequately described in the existing taxonomy (Table 4). Despite having fewer samples, semantically distinct categories like Insurance and Inheritance Procedures perform exceptionally well, showing that precise classification and possibly improved customer outcomes are made possible by distinct service boundaries and specialized terminology.

These findings have significant implications for customer service management and financial operations. Approximately 70% of incoming complaints and requests can be automatically routed to the right specialist team without human intervention according to the system's 70% overall hierarchical accuracy and 83% Level 1 accuracy. Because tickets are instantly routed to agents with the necessary knowledge rather than needing to be manually triaged, this automation feature directly results in lower operating costs and quicker response times. Implementing a confidence threshold strategy, where high-confidence Level 1 predictions move on to automated Level 2 classification while uncertain cases are reviewed by humans, may optimize the balance between automation rate and accuracy, as indicated by the 84% conditional accuracy for Level 2 classifications when Level 1 is correct.

Conclusion and Future Directions

With 83% accuracy at Level 1 and 70% overall hierarchical accuracy across the Aktif Bank and N Kolay datasets, this study shows how successful BERT refined with LoRA is for automated hierarchical classification of financial service

complaints and requests. In terms of true positive rates, the BERT model consistently beat conventional ML techniques by 5–10%, and it performed well in minority classes where ML techniques had trouble. The system's capacity to automatically route roughly 70% of incoming complaints to the right specialized teams marks a significant operational improvement over manual triage systems, even though the inherent class imbalance in complaint data continues to be a fundamental challenge.

Future optimization efforts should concentrate on enhancing initial categorization to achieve the most performance gains, as seen by the conditional accuracy of 84% for Level 2 forecasts when Level 1 classification is accurate. This automated classification system solves important problems in financial service operations while preserving enough accuracy for real-world implementation by facilitating quicker response times, cutting operational expenses, and guaranteeing that complaints are handled by agents with the necessary experience.

There are a number of promising approaches to improving the system's functionality and performance. Implementing hybrid architectures that use computationally efficient standard ML techniques for well-represented classes and BERT for difficult minority classes may offer the best cost-performance trade-off for efficient inference. It may be possible to improve complaint categorization schemas by examining if poor performance in particular categories is due to shortcomings in the current taxonomic structure rather than just model constraints. Together, these improvements could improve automated complaint management's business value and technical capabilities.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPESS journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest.

Funding

* This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Education in Technology and Engineering (www.iconete.net) held in Antalya/Türkiye on November 12-15, 2025.

* We gratefully acknowledge the support and contribution of the Aktif Bank R&D Center throughout the research and analysis process.

References

Abaskohi, A., Rothe, S., & Yaghoobzadeh, Y. (2023). LM-CPPF: Paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning. *arXiv preprint arXiv:2305.18169*.

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. *Expert Systems with Applications, 127,* 256-271.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR, 1*(2), 3.

Kelm, A., Plebański, P., & Kłopotek, R. A. (2024, November). Impact of deep learning-based text feature extraction methods on binary classification quality of customer service call transcripts. In 2024 *IEEE 17th International Scientific Conference on Informatics (Informatics)* (pp. 138-145). IEEE.

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244,* 122778.

Poczeta, K., Płaza, M., Michno, T., Krechowicz, M., & Zawadzki, M. (2023). A multi-label text message classification method designed for applications in call/contact centre systems. *Applied Soft Computing*, *145*, 110562.

Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2025). Think before you classify: The rise of reasoning large language models for consumer complaint detection and classification. *Electronics*, *14*(6), 1070.

Song, W., Rong, W., & Tang, Y. (2024). Quantifying risk of service failure in customer complaints: A textual analysis-based approach. *Advanced Engineering Informatics, 60,* 102377.

Taskiran, S. F., Turkoglu, B., Kaya, E., & Asuroglu, T. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Scientific Reports*, *15*(1), 21631.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., ... & Kenealy, K. (2024). Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295.*

Vairetti, C., Aránguiz, I., Maldonado, S., Karmy, J. P., & Leal, A. (2024). Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making. *European Journal of Operational Research*, *312*(3), 1108-1118.

Xu, W., Yao, Z., Ma, Y., & Li, Z. (2025). Understanding customer complaints from negative online hotel reviews: A BERT-based deep learning approach. *International Journal of Hospitality Management*, *126,* 104057.

Yadav, V., Tang, Z., & Srinivasan, V. (2024). Paraphrase and aggregate with large language models for minimizing intent classification errors*. arXiv preprint arXiv:2406.17163*.

Yilmaz, C., Varnali, K., & Kasnakoglu, B. T. (2016). How do firms benefit from customer complaints?. *Journal of Business Research*, 69(2), 944-955

Zangari, A., Marcuzzo, M., Schiavinato, M., Gasparetto, A., & Albarelli, A. (2023). Ticket automation: An insight into current research with applications to multi-level classification scenarios. *Expert Systems with Applications*, *225,* 119984.

Zhang, D., Mi, R., Zhou, P., Jin, D., Zhang, M., & Song, T. (2024, March). Large model-based data augmentation for imbalanced text classification. In 2024 *5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)* (pp. 1006-1010). IEEE.

Author(s) Information

Simge Senyuz
Aktif Investment Bank Inc.
Esentepe District, Kore Şehitleri Street No: 8/1 Şişli
Istanbul, Turkiye
Contact e-mail: *Simge.Senyuz@aktifbank.com.tr*

Erkut Baloglu
Aktif Investment Bank Inc.
Esentepe District, Kore Şehitleri Street No: 8/1 Şişli
Istanbul, Turkiye

Eren Caglar
Aktif Investment Bank Inc.
Esentepe District, Kore Şehitleri Street No: 8/1 Şişli
Istanbul, Turkiye

Ismail Gocmez
Aktif Investment Bank Inc.
Esentepe District, Kore Şehitleri Street No: 8/1 Şişli
Istanbul, Turkiye

To cite this article:

Senyuz, S., Baloglu, E., Caglar, E., & Gocmez, I. (2025). Classification of user complaints and requests in customer service using Bert Fine-Tuned with Lora. *The Eurasia Proceedings of Educational and Social Sciences (EPESS), 46,* 125-136.