UNIVERSITATEA TEHNICĂ A MOLDOVEI

Technical University of Moldova

EVALUAREA PERFORMANȚEI MODELELOR DE ÎNVĂȚARE AUTOMATĂ PENTRU OPTIMIZAREA SISTEMELOR AUTOMATIZATE DE DIRECȚIONARE A TICHETELOR (DISTILBERT, DISTILROBERTA, XLM-ROBERTA)

PERFORMANCE EVALUATION OF MACHINE LEARNING MODELS FOR AUTOMATED TICKET ROUTING SYSTEMS (DISTILBERT, DISTILROBERTA, XLM-ROBERTA)

Student: gr. IS-231M, Dumbrăveanu Ecaterina

Coordinator: Beşliu Corina, university lecturer

Chisinau, 2025

MINISTRY OF EDUCATION AND RESEARCH OF THE REPUBLIC OF MOLDOVA

Technical University of Moldova

Faculty of Computers, Informatics, and Microelectronics

Department of Software Engineering and Automation

Approved for defense
Department head:
Ion FIODOROV, phd, associate professor

„___” ________ 2025

PERFORMANCE EVALUATION OF MACHINE LEARNING MODELS FOR AUTOMATED TICKET ROUTING SYSTEMS (DISTILBERT, DISTILROBERTA, XLM-ROBERTA)

Master's project

Student:	______________	**Dumbrăveanu Ecaterina, IS-231M**
Coordinator:	______________	**Beşliu Corina, university lecturer**
Consultant:	______________	**Cojocaru Svetlana, university assistant**

Chisinau, 2025

ABSTRACT

This thesis focuses on the review and performance testing of machine learning models for automated ticket routing systems, an essential instrument for raising customer satisfaction and operational effectiveness in IT support and customer service applications. In order to automate ticket routing, the project starts with a domain analysis and identification of the primary issues with such systems that come up. Machine learning and natural language processing (NLP) have been highlighted as being crucial to this process.

The Large Language Models (LLMs) that have transformed NLP jobs are extensively covered at the beginning of the second chapter, which looks into the machine learning models used for ticket classification. It explains some BERT based models and techniques to fine-tune them, in extensive detail and shows how they may be used to improve ticket routing systems' accuracy. The thesis also discusses the class imbalance problem, which is a prevalent problem in ticket routing when certain ticket categories are dominated by some other classes.

Throughout the thesis, various machine learning models like DistilBERT, DistilRoBERTa, and XLM-RoBERTa are tested and compared using real-world datasets, and methods for increasing accuracy and efficiency are investigated. The study ends with suggestions for putting into practice a reliable, machine-learning-based system that can route tickets automatically and effectively. Its efficacy is demonstrated by benchmarking the system's results against traditional approaches. The best results for the automated ticket routing classification task were obtained for the DistilBERT base uncased model with batch size equal to 8 and trained on 3 epochs.

REZUMAT

Această teză se concentrează pe revizuirea și testarea performanței modelelor de învățare automată pentru sistemele automate de direcționare a tichetelor, un instrument esențial pentru creșterea satisfacției clienților și a eficienței operaționale în aplicațiile de asistență IT și de servicii pentru clienți. Pentru a automatiza direcționarea tichetelor, proiectul începe cu o analiză a domeniului și identificarea problemelor primare care pot apărea în astfel de sisteme. Învățarea automată și procesarea limbajului natural au fost evidențiate ca fiind esențiale pentru acest proces.

Modelele preantrenate care au transformat principiile de muncă a procesării limbajului natural sunt discutate pe larg la începutul celui de-al doilea capitol, care analizează modelele de învățare automată utilizate pentru clasificarea tichetelor. Aceasta explică în detaliu modele bazate pe BERT la fel și tehnici de fine-tuning și arată cum acestea pot fi utilizate pentru a îmbunătăți acuratețea sistemelor de direcționare a tichetelor. Teza discută, de asemenea, problema dezechilibrului în clase, care este o problemă predominantă în direcționarea a tichetelor atunci când anumite categorii de tichete sunt dominate de alte clase, și prezintă modele preantrenate cu un număr mai mic de parametri, demonstrând eficacitatea lor în contexte limitate de resurse.

La fel, sunt testate și comparate diferite modele de învățare automată, precum DistilBERT, DistilRoBERTa și XLM-RoBERTa folosind seturi de date din aplicații reale și sunt investigate metode pentru creșterea preciziei și eficienței. Studiul se încheie cu sugestii pentru punerea în practică a unui sistem de eficient, care poate ruta tichetele automat și eficient. Eficacitatea sa este demonstrată prin compararea rezultatelor sistemului față de abordările tradiționale. Cele mai bune rezultate pentru sarcina de rutare automată a biletelor au fost obținute pentru modelul DistilBERT cu dimensiunea lotului egală cu 8 și antrenat pe 3 epoci.

CONTENTS

INTRODUCTION **9**
1 DOMAIN ANALYSIS AND ARGUMENTATION **11**
1.1 Problem definition **11**
1.2 Machine learning models in NLP tasks 12
1.3 Goals and objectives 13
2 MACHINE LEARNING (ML) MODELS **15**
2.1 Large Language Models (LLM) 15
2.1.1 BERT based models 16
2.1.2 Fine-tuning 17
2.2 Small Large Language Models (SLMs) 18
2.3 Class imbalance problem 20
2.4 Models selecting approach 21
3 DATA ANALYSIS AND PREPARATION FOR TRAINING **25**
3.1 Dataset description 25
3.2 Dataset fields 26
3.3 Exploratory data analysis 30
3.4 Data cleaning and augmentation 36
4 TESTED MODELS **40**
4.1 DistilBERT base model (uncased) 40
4.2 DistilRoBERTa base model 43
4.2 XLM-RoBERTa (base-sized model) 45
5 STRATEGIES FOR IMPROVING MODEL PERFORMANCE **48**
5.1 Stop-words removal 48
5.2 Tokenization approach 49
5.3 Batch size parameter adjustment 53
5.4 Loss function adjustment of the classification model 55
5.5 Number of epochs adjustment 56
CONCLUSIONS **59**
BIBLIOGRAPHY **61**

INTRODUCTION

In the era of digital transformation and the evolution of all commercial domains, the most successful companies are those that interact with their customers, try to stay tuned with their domains' evolution, and change their business to provide good service and satisfy their customers. Customer feedback is important in developing weak points of the company and pointing out areas in which services, products, and the general consumer experience can all be improved. Customer satisfaction is not only a goal but also a need in today's highly saturated industry to sustain long-term success. Businesses constantly looking out and responding to customer feedback are more qualified to solve customer concerns, develop new ideas, and adjust to changes in the market. Additionally, customer service now involves more than just solving a single client's issue quickly. It's about establishing a long-term partnership in which every consumer interaction presents chances for a deeper engagement.

Also, expectations from consumers are rising. Consumers are prepared to find new service providers if they do not get supported with the prompt, simple, and efficient service they expect. Moreover, a company's ability to meet or beyond client expectations is frequently correlated with its performance in the marketplace.

Efficient handling of client inquiries, complaints, and service requests is a crucial component in preserving successful customer relationships. On-time resolution of customer issues and customer satisfaction depend on the ability to process and respond to them accurately and quickly. But when companies get bigger, the number of client tickets increases as well, which makes it harder for human processes to keep up. Automated solutions that can effectively handle and route tickets to the right teams or departments are therefore extremely important.

Automated ticket routing systems, provided by machine learning models are good solutions to handle customer requests. Such systems improve the accuracy and speed of ticket processing while also reducing the workload for human operators. Such systems improve the accuracy and speed of ticket processing while also reducing the workload for human operators. Machine learning models are able to predict correctly which team or individual will be most qualified to handle a particular ticket by using powerful algorithms and previous data. This ensures that issues will be resolved more quickly and efficiently. However, the performance of the machine learning models that power these automated systems has a significant impact on how effective they are. Models with low performance might cause delays, tickets routed to the wrong teams, and dissatisfied customers. To guarantee that these models meet the performance requirements necessary in a dynamic business environment, it is important to evaluate and optimize them.

Since most support tickets are composed of unstructured text that needs to be understood and categorized, natural language processing, or NLP, is crucial for automated ticket routing systems. Trained on past ticket data, machine learning models are able to recognize similarities and trends between related tickets, which enables them to forecast the best possible resolution path. In addition to greatly increasing the efficiency and speed of ticket handling, this automation reduces the workload for human operators and guarantees that problems are resolved quickly.

The process of automatically routing tickets is made more difficult by the changing nature of customer support configurations. Models may need to adjust to new kinds of problems and adjust ticket patterns without continuous retraining. In sectors such as IT service management, where new software updates or hardware faults can bring new types of tickets, such situations call for a strong system that can handle unseen or rarely encountered categories. These conditions can be challenging for machine learning models that are trained only using past data without adaptation strategies. This emphasizes the need for more adaptable learning strategies, which enables models to generalize to new tasks with little further training.

Managing class imbalance is another crucial factor in the context of automated ticket routing. In many service environments, certain types of requests are far more common than others and may dominate the ticket volume, while more complex, high-priority issues occur far less frequently. This imbalance may cause a machine learning model to disproportionately benefit the majority class, misclassifying minority classes, which are frequently those with more complicated and important problems. Consequently, in order to guarantee that the automated routing system functions effectively for all ticket types, not just the most common ones, it is crucial to address this class imbalance issue.

Extensive evaluation and optimization are needed to guarantee that machine learning models in these systems work as intended. In order to determine whether the model is suitable for use in real-world scenarios, key performance factors like accuracy, precision, and the capacity to generalize across various ticket categories must be analyzed. Other essential considerations in evaluating the model's success are its capacity to manage unstructured data, its scalability, and its flexibility in responding to changing client demands. Furthermore, the model's latency (how quickly it processes and routes tickets) must also be taken into account in scenarios where prompt responses are essential.

BIBLIOGRAPHY

1. L. Kaňovská, “Customers service and their role for industrial small and medium companies,” vol. 14, Jan. 2009.

2. N. Arici, L. Putelli, A. E. Gerevini, L. Sigalini, and I. Serina, “LLM-based Approaches for Automatic Ticket Assignment: A Real-world Italian Application,” in *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023)*, E. Bassignana, D. Brunato, M. Polignano, and A. Ramponi, Eds., in CEUR Workshop Proceedings, vol. 3551. Rome, Italy: CEUR, Nov. 2023.

3. P. Johri, S. K. Khatri, A. Al-Taani, M. Sabharwal, S. Suvanov, and A. Chauhan, “Natural Language Processing: History, Evolution, Application, and Future Work,” 2021.

4. M. U. Hadi *et al.*, *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects*. 2023.

5. Y. Wang, W. Qu, and X. Ye, “Selecting Between BERT and GPT for Text Classification in Political Science Research,” Nov. 07, 2024, *arXiv*: arXiv:2411.05050. doi: 10.48550/arXiv.2411.05050.

6. A. Rogers, O. Kovaleva, and A. Rumshisky, “A Primer in BERTology: What we know about how BERT works,” Nov. 09, 2020, *arXiv*: arXiv:2002.12327. doi: 10.48550/arXiv.2002.12327.

7. K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text Classification Algorithms: A Survey,” Information, vol. 10, no. 4, p. 150, Apr. 2019.

8. J. T. Behrens, “Principles and procedures of exploratory data analysis.,” *Psychological Methods*, vol. 2, no. 2, pp. 131–160, 1997.

9. M. Komorowski, D. C. Marshall, J. D. Salciccioli, Y. Crutain .“Exploratory Data Analysis”, 2016.

10. G. M. Weiss and F. Provost, “The effect of class distribution on classifier learning: an empirical study,” Rutgers University, 2001.

11. S. Y. Feng *et al.*, “A Survey of Data Augmentation Approaches for NLP,” Dec. 01, 2021, *arXiv*: arXiv:2105.03075.

12. S. ES, “Data Augmentation in NLP: Best Practices From a Kaggle Master,” neptune.ai. Available: https://neptune.ai/blog/data-augmentation-nlp

13. K. Zhang *et al.*, “Fast and Slow Generating: An Empirical Study on Large and Small Language Models Collaborative Decoding,” Oct. 23, 2024.

14. “Hugging Face – The AI community building the future.” [Online]. Available: https://huggingface.co/

15. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” Mar. 01, 2020.

16. “distilbert/distilbert-base-uncased · Hugging Face.” Available: https://huggingface.co/distilbert/distilbert-base-uncased.

17. “distilbert/distilroberta-base · Hugging Face.” Available: https://huggingface.co/distilbert/distilroberta-base

18. “FacebookAI/xlm-roberta-base · Hugging Face.” Available: https://huggingface.co/FacebookAI/xlm-roberta-base

19. F. Rodriguez, “An Introduction to Information Retrieval”. Available: https://www.academia.edu/27076940/An_Introduction_to_Information_Retrieval

20. J. L. Gastaldi, J. Terilla, L. Malagutti, B. DuSell, T. Vieira, and R. Cotterell, “The Foundations of Tokenization: Statistical and Computational Concerns,” Nov. 04, 2024, *arXiv*: arXiv:2407.11606. doi: 10.48550/arXiv.2407.11606.

21. À. Atrio and A. Popescu-Belis, “Small Batch Sizes Improve Training of Low-Resource Neural MT,” in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, S. Bandyopadhyay, S. L. Devi, and P. Bhattacharyya, Eds., National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI), Dec. 2021, pp. 18–24. Available: https://aclanthology.org/2021.icon-main.4

22. “Loss Functions in Machine Learning Explained.”
Available: ttps://www.datacamp.com/tutorial/loss-function-in-machine-learning

23. B. M. Hussein, S. M. Shareef, “An Empirical Study on the Correlation between Early Stopping Patience and Epochs in Deep Learning,” *ResearchGate*, Oct. 2024, doi: 10.1051/itmconf/20246401003.