DistilBERT-based Classification of Hungarian Banking Complaints under Regulatory Constraints

Zsolt Krutilla[1,2]

[1] Obuda University, John von Neumann Faculty of Informatics,
Becsi ut 96/b, 1034 Budapest, Hungary
E-mail: krutilla.zsolt@nik.uni-obuda.hu

[2] University of Dunaujvaros, Institute of Computer Engineering, Department of Software Development and Application
Tancsics M. 1/A, 2400 Dunaujvaros, Hungary
E-mail: krutillazs@uniduna.hu

Abstract: The purpose of the present research is to develop a natural language processing (NLP) model that can automatically categorize Hungarian-language bank complaint reports according to the categories defined in the Hungarian National Bank's Decree 57/2023 (XI. 24.). Following the successful implementation of the initial dictionary-based model, the focus of the research shifted to the development and validation of an artificial intelligence-based, fine-tuned DistilBERT model, which offers a more efficient, faster and economical solution for the automated processing of banking customer communications. During the learning process, significant accuracy gains were achieved by optimized batch and epoch parameterization, while taking into account sustainability and cost-effectiveness aspects. The novelty of this research lies in the creation of a unique NLP model optimized for bank texts in Hungarian, based on large-scale real-world data relevant to the financial sector. The results contribute to the digitization of bank customer service, the acceleration of complaint handling processes and the improvement of risk management efficiency.

Keywords: natural language processing; distilbert; banking environment; complaint handling

1 Introduction

Natural Language Processing (NLP) has evolved significantly over the past decades, enabling computers to understand and process human language more and more efficiently. A wide range of NLP applications include text categorization, machine translation, sentiment analysis, and automated customer service [1], [2]. In particular, the use of NLP technologies is gaining momentum in the financial sector,

where the rapid and accurate processing of large amounts of unstructured text data is crucial for effective risk management and customer care [3], [4].

In the field of banking complaint handling, natural language processing offers the possibility to automatically categorize and prioritize incoming complaints, thus ensuring faster and more efficient responses [5]. However, the specificity of the Hungarian language, such as the complexity of its rich suffixation and syntactic structures, poses a challenge for the development of NLP models, especially for deep learning-based solutions [6].

Traditional dictionary-based and rule-based methods have the advantage of transparency and ease of implementation, but they are limited in their applicability to large-volume, diverse texts and do not handle linguistic nuances and synonyms well [7]. However, in recent years, transformer-based models such as BERT and its lightweight variants (DistilBERT, TinyBERT) have made significant breakthroughs in language modelling due to their ability to learn context-dependent representations, improving the accuracy and generalizability of categorization [8], [9]. In particular, fine-tuned domain-specific models have proven to be effective in the processing of financial texts [10].

However, automated systems for processing bank complaints in Hungarian are less widespread in the international literature, so the novelty of the present research is that it develops an efficient and sustainable NLP model tailored to the category system defined by the "Magyar Nemzeti Bank" (Hungarian National Bank), which operates on a real dataset of bank complaints. According to the MNB Regulation 57/2023 (24.XI.), the professional and rapid categorization of complaints is not only a legal requirement, but also a key element of financial market stability [11].

Therefore, in the first phase of the research, a dictionary-based model was developed, which worked well, but had scalability and flexibility limitations. This was further developed by fine-tuning a DistilBERT-based model that was able to achieve significant accuracy gains while also reducing the computational and environmental burden [12]. In parallel, research has focused on the determination of optimal training parameters such as batch size and epoch number, which are critical for the efficiency and sustainability of the model [13].

2 Research background and approach

Reliable and scalable text classification constitutes a core component in the automation of banking complaint management systems. In the initial phase of the present research, a dictionary-based text classification approach was applied, in which manually curated keyword sets were associated with predefined complaint categories. Formally, let $D = \{d_1, d_2, \dots, d_K\}$ denote the set of category-specific

dictionaries, where each d_k contains a finite set of keywords. For a given complaint text x, the assigned category $\hat{y}$ is determined as

$$\hat{y} = \arg \max_{k \in \{1,\ldots,K\}} \sum_{w \in x} \mathbf{1}(w \in d_k), \quad 2.1$$

where $\mathbf{1}(\cdot)$ denotes the indicator function.

This rule-based strategy demonstrated high classification accuracy, particularly in the case of well-structured and linguistically formal complaint texts, where lexical cues are explicit and unambiguous [14].

Despite these favorable results, dictionary-based classification systems suffer from inherent structural limitations. Most notably, their performance critically depends on continuous human intervention, as the underlying keyword inventories must be manually updated to accommodate emerging linguistic patterns, newly introduced terminology, semantic drift, or previously unseen complaint topics. This maintenance requirement results in significant operational overhead and restricts the long-term scalability and adaptability of the approach [15].

Furthermore, dictionary-based methods exhibit limited capacity for contextual interpretation, as they primarily rely on surface-level lexical matching rather than semantic representation. Consequently, they are unable to robustly capture syntactic dependencies, word order, or implicit meaning conveyed through context. This limitation is particularly pronounced in the Hungarian language, which is characterized by rich morphology, free word order, and extensive agglutination, all of which substantially increase lexical variability and reduce the effectiveness of static keyword-based matching strategies [16].

To overcome these problems, the next stage of the research turned its attention to language models based on artificial intelligence. In particular, I investigated the so-called pretrained models, which have already proven their effectiveness in various natural language processing (NLP) tasks. In the course of the research, I aimed at training and evaluating these models on the same annotated Hungarian complaint dataset that I have previously used for the dictionary-based approach (objective performance comparison).

Among the models tested were the Hungarian-optimized version of BERT (HuBERT [17]), the generative GPT-3 [18], and the Hungarian-language-specific PULI-GPT model developed in-house [19]. In addition, various fine-tuning techniques and deep learning algorithms were experimented with, such as the Longformer and RoBERTa adaptations, which offer advantages in dealing with longer texts [20].

The training of the models was made possible by a self-invested computer system optimised specifically for deep learning tasks. The system is equipped with an NVIDIA RTX 4090 GPU, AMD Ryzen 7 3800X processor and 36 GB of RAM to teach and test the computationally demanding models. The training and fine-tuning processes were executed in both TensorFlow and PyTorch environments to ensure

compatibility, reproducibility, and optimal performance across different deep learning frameworks. Leveraging the respective strengths of these platforms enabled efficient experimentation with advanced optimization strategies and facilitated seamless integration of pre-trained models and tokenizers into the fine-tuning pipeline. This infrastructure has significantly reduced the learning time and allowed us to experimentally run more complex models.

However, the use of advanced language models poses challenges not only from a technological but also from an environmental point of view. The energy requirements for training and the resource consumption of longer run-time models can leave a significant ecological footprint [21]. With this in mind, research has focused on maximizing efficiency, optimizing model sizes, and measuring inference time and energy consumption to ensure that the developed system remains sustainable and economical in the long-term.

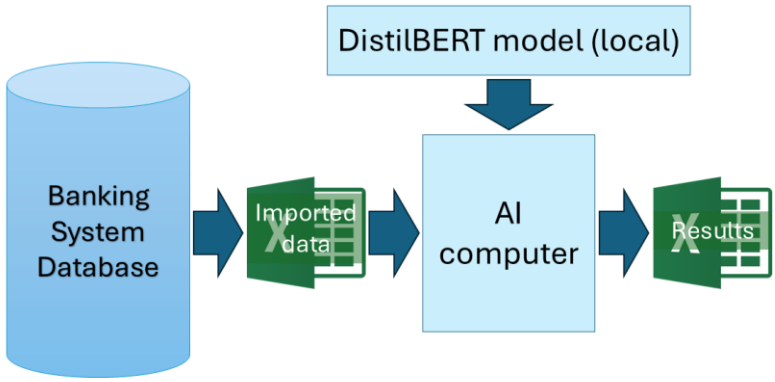

Figure 1 – Schematic structure of the system

2.2 Model choice and rationale: the use of DistilBERT in bank complaint handling

Following the successful implementation and validation of a rule-based and dictionary-based text classification approach in the first phase of the research, I extended the investigation toward artificial intelligence–based methods. The objective of this extension was to complement the earlier solution – characterized by substantial manual maintenance requirements – with trainable language models capable of learning deeper contextual representations and supporting long-term scalability. Accordingly, in the second phase of the research, I focused on identifying language models that are well suited for text classification tasks, with particular emphasis on the automatic categorization of banking complaint submissions, drawing on findings from both domestic and international scientific literature.

Based on the literature review, models built upon the BERT (Bidirectional Encoder Representations from Transformers) architecture proved to be the most promising, as they demonstrate outstanding performance in learning context-dependent language representations and in text understanding tasks [14], [17].

The present research is not intended to provide an exclusive evaluation of a single model architecture; rather, it aims to establish a progressively extensible experimental framework that enables the comparative analysis of language models with different capacities and language-specific adaptations. Accordingly, in later stages of the research, I plan to extend the investigation to transformer-based architectures explicitly optimized for the Hungarian language, such as HuBERT, as well as to higher-capacity models with a larger number of parameters, including PULI-BERT-Large.

In the present study, however, the DistilBERT model was selected and introduced as the initial reference architecture, as it offers a favorable trade-off between linguistic performance, computational resource requirements, and inference efficiency. The use of DistilBERT enables the establishment of a stable and reproducible baseline against which the performance of larger-capacity or language-specifically fine-tuned models can be quantitatively compared in future work.

The model selection was driven not only by linguistic performance but also by considerations of computational cost-efficiency and environmental sustainability. In the design of responsible artificial intelligence applications, the optimization of available computational resources and the minimization of energy consumption have become increasingly important requirements [21]. Accordingly, in this research I adopted DistilBERT, a compressed variant of the BERT architecture, which – through the knowledge distillation process – achieves substantial resource savings while preserving most of the linguistic capabilities of the full model [9].

The number of parameters in DistilBERT is approximately 40% of that of the full BERT model, i.e.,

$$| \theta_{\text{DistilBERT}} | \approx 0.4 \cdot | \theta_{\text{BERT}} |,$$

while its performance on benchmark tasks reaches approximately 97% of the accuracy of the full model [9]. The reduction in parameter count has a direct impact on the computational complexity required during inference, which, for transformer-based architectures, is well approximated as being proportional to the number of model parameters. Consequently, the reduction in inference time per document can be approximated as

$$t_{\text{inf}}^{\text{DistilBERT}} \approx \alpha \cdot t_{\text{inf}}^{\text{BERT}}, \alpha \approx 0.4\text{-}0.5,$$

which is consistent with empirical observations indicating that the inference time of DistilBERT is approximately 60% lower than that of the full BERT model [9].

This property is particularly relevant in the automation of banking complaint handling systems, where response time, scalability, and operational stability constitute critical performance factors. If a system processes N complaint texts on a daily basis, the total inference time can be well approximated by the following relation:

$$t_{\text{inf,total}} \approx N \cdot t_{\text{inf,doc}}.$$

Within this framework, reducing the inference time per document linearly decreases the overall computational demand, which directly contributes to lower energy consumption and reduced operational costs.

The DistilBERT architecture effectively exploits the high degree of parallel computational capacity offered by modern GPU architectures – such as the NVIDIA RTX 4090. Owing to the high FP16 and Tensor Core performance of the RTX 4090, the compressed model can be executed with high throughput, while its reduced memory footprint allows for larger batch sizes and more efficient utilization of GPU memory. As a result, the energy consumption per processed document is reduced, which is consistent with the fundamental principles of designing sustainable and cost-efficient artificial intelligence systems.

Preliminary pilot experiments were conducted with larger and Hungarian-specific transformer architectures, including HuBERT and PULI-GPT, on the same annotated complaint dataset. These exploratory runs indicated comparable or slightly higher classification performance in certain configurations; however, this came at the cost of substantially increased training time, GPU memory usage, and energy consumption. Given the applied focus of the present study and the emphasis on efficiency, reproducibility, and sustainability, DistilBERT was selected as the primary reference architecture. A systematic, quantitative comparison with larger and language-specific models is planned as part of future work.

2.3 Data preparation and learning methodology

The training phase was based on a dataset consisting of real banking customer complaints that had been manually annotated in a prior process and made available in an anonymized form by the bank. The anonymization procedures were designed to protect both customer and institutional data, in compliance with applicable data protection and ethical regulations (e.g., removal of personal identifiers, masking of numerical data, and obfuscation of account numbers). The reliability and consistency of the dataset were ensured through the controlled origin of the data and a multi-level expert annotation process governed by an internal procedural framework.

The annotation was conducted within a product-category-based organizational structure aligned with banking operations. Complaint submissions were first assigned to the relevant banking product group (e.g., retail current account), followed by the identification of the appropriate thematic category based on the

content of the complaint (e.g., billing-related issues). Expert annotation teams associated with each product category typically consisted of 5-10 members, and annotation decisions were made according to a four-eyes principle, with each decision subsequently reviewed and approved by a group leader.

The professional interpretation, categorization of complaints, and preparation of response letters were carried out by dedicated complaint-handling experts, while final approval and authorization were the responsibility of group leaders. Eligibility for the expert role required a minimum of five years of relevant front line banking experience, whereas group leader positions were typically held by staff members with 15-20 years of professional experience. This competence-based division of responsibilities and control structure was formalized through an internal banking procedure, which provided a uniform framework to ensure the consistent application of annotation decisions.

As a result, the annotated dataset contained high-quality reference labels suitable for the reliable training of supervised learning algorithms. Nevertheless, additional preprocessing and cleaning of the raw textual data were required during the preparation of the model training process, with particular attention to handling formatting issues and linguistic inconsistencies.

One of the primary challenges during preprocessing was the presence of character encoding inconsistencies. The language model exclusively supports UTF-8 encoding; however, the source files contained non-standard or unrecognizable special characters, including malformed diacritics, characters from non-Latin writing systems, and control sequences. The removal and normalization of these elements constituted a critical step in ensuring a stable and noise-free training process. Additional cleaning steps included lower casing, the removal of special characters, and the reduction of textual noise.

Model evaluation was performed on a strictly controlled test dataset that was fully separated from the training corpus and consisted exclusively of a disjoint subset of the previously annotated data. The datasets used during the training and testing phases originated from two independent sets of complaint letters; therefore, individual documents did not appear simultaneously in either the training or evaluation stages. The size of the training dataset was $N_{\text{train}} = 128\,930$, while the size of the test dataset was $N_{\text{test}} = 39\,421$.

The evaluation was based on so-called gold standard categories defined by independent domain experts. This approach provided an objective foundation for assessing the model's performance using business-relevant metrics, such as accuracy, the F1-score, and class-wise precision.

The carefully structured data processing and training strategy established a solid methodological foundation for the model's subsequent applicability in real-world business environments. Training and evaluation conducted on strictly separated

datasets contributed to ensuring that the obtained results remained both scientifically reliable and interpretable from a business perspective.

Model evaluation was carried out along multiple complementary performance metrics in order to quantitatively assess different aspects of classification performance. The evaluation primarily relied on three fundamental metrics: precision, recall, and their harmonic mean, the F1-score.

Precision expresses the proportion of cases labeled as positive by the model that are in fact correctly classified. Formally, precision is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2.2}$$

where TP (True Positives) denotes the number of correctly classified positive instances, and FP (False Positives) denotes the number of incorrectly classified positive instances.

Recall measures the proportion of actual positive cases that are correctly identified by the model. Its mathematical formulation is given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.3}$$

where FN (False Negatives) denotes the number of positive instances incorrectly classified as negative.

The F1-score is the harmonic means of precision and recall, providing a balanced metric in scenarios where it is necessary to evaluate the trade-off between these two measures:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \tag{2.4}$$

The above metrics were computed separately for each class, taking into account the substantial variation in sample sizes across categories within the training dataset (a total of 31 categories). Consequently, relying solely on aggregated performance metrics would yield a distorted representation of the model's true classification performance.

To characterize the overall classification performance, the accuracy metric was computed, which expresses the proportion of all correctly predicted instances relative to the total number of samples:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2.4}$$

where TN (True Negatives) denotes the number of correctly classified negative instances.

In order to ensure comparability of performance across categories, macro-averaged metrics were also computed. These metrics represent the simple arithmetic mean of

the class-specific measures, independently of the number of samples in each class. The definitions of macro-averaged precision, recall, and F1-score are as follows:

$$\text{Macro Precision} = \frac{1}{C}\sum_{c=1}^{C} \text{Precision}_c \tag{2.5}$$

$$\text{Macro Recall} = \frac{1}{C}\sum_{c=1}^{C} \text{Recall}_c \tag{2.6}$$

$$\text{Macro F1-score} = \frac{1}{C}\sum_{c=1}^{C} \text{F1-score}_c \tag{2.7}$$

where $C = 31$ denotes the number of classes.

To account for the effect of differences in sample sizes across classes, weighted average metrics were also applied, in which the contribution of each class-specific metric is proportional to the number of samples in the given class. The weighted precision, recall, and F1-score are computed as follows:

$$\text{Weighted Precision} = \frac{\sum_{c=1}^{C} \text{Precision}_c \cdot n_c}{\sum_{c=1}^{C} n_c} \tag{2.8}$$

$$\text{Weighted Recall} = \frac{\sum_{c=1}^{C} \text{Recall}_c \cdot n_c}{\sum_{c=1}^{C} n_c} \tag{2.9}$$

$$\text{Weighted F1-score} = \frac{\sum_{c=1}^{C} \text{F1-score}_c \cdot n_c}{\sum_{c=1}^{C} n_c} \tag{2.10}$$

where n_c denotes the number of samples belonging to the c-th class.

2.4 Input data distribution

The performance of supervised learning-based text classification models is fundamentally determined by the class distribution of the training dataset. In the present research, the training corpus was designed for the categorization of banking complaint letters, and preliminary data analysis revealed a pronounced imbalance in the class distribution. The number of samples associated with individual categories spans several orders of magnitude, which makes the classification task particularly challenging from a methodological perspective.

The categories with the largest number of samples (top 6 categories in Fig. 2) account for a substantial proportion of the overall dataset. In contrast, several categories are represented by only a very small number of instances, including highly specific complaint types such as *"jogosulatlan tevekenyseg"* (4 samples), *"elozetes tajekoztatasi kotelezettseg"* (1 sample), and *"arubemutatoval egybekotott termekertekesites soran nyujtott penzugyi szolgaltatas"* (1 sample).

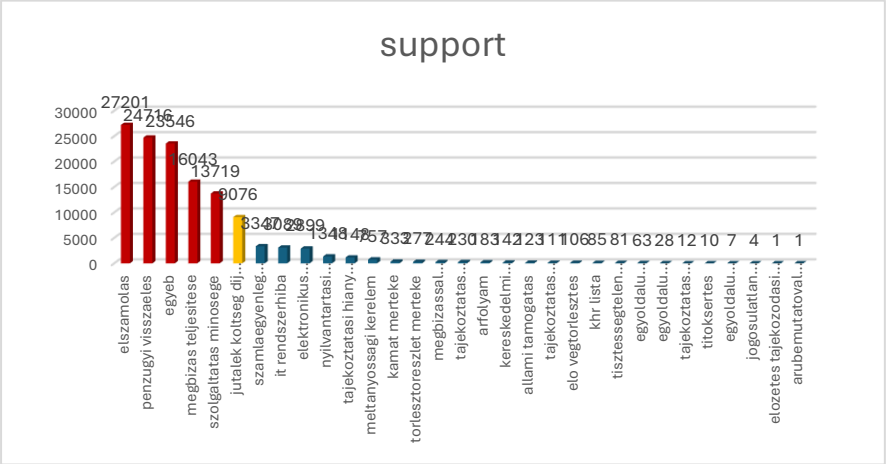

Figure 2 – Input data distribution by categories

Formally, the degree of class distribution skewness can be characterized by the ratio of the maximum to the minimum class support:

$$\frac{\max\limits_{c} n_c}{\min\limits_{c} n_c} \gg 1,$$

where n_c denotes the number of samples belonging to the c-th category. In the case of the present dataset, this ratio spans several orders of magnitude, which clearly indicates a strongly imbalanced training corpus.

A direct consequence of such a distribution is that global performance metrics – such as aggregated accuracy – do not, in isolation, provide an adequate representation of the model's true behavior, as the dominance of high-frequency categories may obscure classification errors associated with rare classes. Accordingly, in the present research, particular emphasis was placed on class-wise performance analysis, as well as on the use of macro-averaged and weighted performance metrics.

Class imbalance is relevant not only from an evaluation perspective but also with respect to the learning process itself, as it influences the optimization of the loss function and the formation of decision boundaries learned by the model. This further justifies the careful selection of hyperparameters and a cautious interpretation of results when working with a heterogeneous dataset that represents a real-world business environment.

2.5 Model parameterization and validation strategy

Following the prior definition of the evaluation metrics (accuracy, F1-score, recall, and precision), I proceeded with fine-tuning the model, placing particular emphasis on the optimization of key hyperparameters that directly influence the dynamics of

the learning process. In the initial configuration, the maximum length of the input sequences was fixed at $L = 512$tokens, which corresponds to the upper limit of the context window supported by transformer-based architectures and provides an appropriate trade-off between contextual coverage and computational requirements.

During optimization, the initial learning rate was set to

$$\eta = 3 \times 10^{-5}$$

and weight updates were performed using the Adam optimization algorithm, implemented within the Keras framework [22]. In the initial phase of training, no early stopping (early escape) mechanism was applied, as the model's behavior and convergence dynamics on the previously unexplored dataset were not known as a priori. Consequently, observing the full training trajectory was methodologically justified in order to empirically analyze the evolution of the loss function, convergence stability, and potential overfitting phenomena.

The Adam algorithm updates model parameters through adaptive estimates of the first- and second-order moments of the gradients:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \tag{2.11}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \tag{2.12}$$

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}, \tag{2.13}$$

where g_t denotes the current gradient, m_t and v_t represent the bias-corrected moment estimates, and β_1, β_2, and ε are hyperparameters controlling the numerical stability of the algorithm.

During the exploration of the hyperparameter space, the size of the training dataset and the available computational resources – particularly GPU memory constraints and processing throughput – defined the feasible range of batch sizes and epoch counts that could be reasonably evaluated. Accordingly, the experimental setup focused on systematically scaling the batch size and the number of training epochs, while all other parameters were kept constant.

Model performance was evaluated through a structured, multi-stage training and feedback process. Multiple training runs were executed in which only the batch size and the epoch count were varied. This approach proved suitable for the detailed analysis of learning curve behavior, convergence speed, and overfitting risk, and it provided a reliable basis for comparing different training configurations.

Particular emphasis was placed on the strict separation of the datasets. The training and validation datasets were constructed in a fully independent manner to ensure that the evaluation of the model's generalization capability remained unbiased. Data partitioning was performed randomly while maintaining statistical balance, thereby ensuring that all classes were represented in both subsets.

3 Results

In addition to the training parameters detailed in *Table 1*, in particular the optimized settings for batch size and epoch number, the model performance was evaluated on a pre-separated validation dataset not seen during training. The results obtained are presented below, allowing a comparison of the different configurations and an objective assessment of the predictive capacity and generalization ability of the model.

The following batch and epoch sizes have been used in the training (Table 1), where the b = batch size and e = epoch number (e.g. b6e8 = 8 batch and 6 epoch):

Table 1 – Settings of epoch connected to batch sizes and its results

Model	W-Precision	W-Recall	W-F1-score	Overall
b6e8	0,65764048	0,64777657	0,62963831	0,64777657
b6e16	0,65975099	0,65982598	0,65732458	0,65982598
b8e8	0,62942455	0,65503158	0,63795215	0,65503158
b10e8	0,64690474	0,65076989	0,63549775	0,65076989
b12e8	0,65781988	0,66299688	0,65280749	0,66299688
b14e8	0,67425667	0,66888207	0,64918656	0,66888207
b16e8	0,64249388	0,64085132	0,63922079	0,64085132
b16e16	0,66024655	0,66358033	0,65764100	0,66358033
b16e32	0,65432066	0,65229193	0,64714427	0,65229193
b18e8	0,65538441	0,65820248	0,65053456	0,65820248
b20e8	0,66486196	0,66413840	0,65131276	0,66413840
b22e8	0,65178274	0,65439740	0,64997444	0,65439740
b26e8	0,66200857	0,66502626	0,65958594	0,66502626
b28e8	0,65721868	0,66928794	0,66039610	0,66928794
b28e16	0,66468393	0,66870450	0,66556928	0,66870450
b28e32	0,66774881	0,66748687	0,66526712	0,66748687
b30e8	0,66421028	0,66157632	0,65764596	0,66157632
b32e8	0,65955886	0,66527993	0,65780898	0,66527993
b34e8	0,66370320	0,66680196	0,66166635	0,66680196
b36e8	0,66189774	0,67101291	0,66131916	0,67101291
b38e8	0,65638410	0,66172852	0,65569143	0,66172852
b38e16	0,66625338	0,67149489	0,66512744	0,67149489
b38e32	0,67308397	0,67920651	0,67329451	0,67920651
b40e8	0,65717388	0,66058700	0,64792376	0,66058700
b42e8	0,66626242	0,65843079	0,65824062	0,65843079
b42e60	0,68309556	0,68072855	0,67754590	0,68072855
b42e100	0,67255342	0,67953629	0,67430552	0,67953629

In the financial sector, and particularly in banking complaint handling environments, the interpretation and weighting of predictive performance metrics differ from common practice in conventional natural language processing applications. Within this context, weighted precision (Weighted Precision) and the weighted F1-score (Weighted F1-score) can be regarded as the primary evaluation criteria, as these metrics provide a realistic assessment of the reliability of the model's predictions while accounting for the unequal support of individual classes.

In contrast, recall or aggregated accuracy, when considered in isolation, have limited interpretative value, as classification correctness is of paramount importance in banking complaint processing. A misclassified complaint – such as a case incorrectly categorized in supervisory or regulatory reports – may have more severe consequences than a submission that is not automatically classified by the model. In the latter case, such complaints are handled as part of standard business processes through manual exception-handling procedures, thereby ensuring the complete and compliant processing of all customer complaints.

During model evaluation, confidence intervals were not computed, as the business-specific characteristics and variability of the application environment do not allow for the definition of stable utility thresholds to which statistically interpretable intervals could be meaningfully assigned. The business objective of the model is not to achieve a formally predefined accuracy level, but rather to reduce human resource requirements through automated pre-screening. The magnitude of this effect cannot be quantified in an exact manner due to the complexity of banking operations and the heterogeneity of complaint handling processes.

Although in theoretical terms, achieving 100% accuracy and recall may be considered an ideal objective, this is not a realistic expectation in a real-world business environment operating on a heterogeneous complaint dataset. Accordingly, the evaluation strategy adopted in the present research focuses on minimizing incorrect automatic classifications and prioritizing high-confidence predictions, in alignment with the operational and regulatory requirements of banking complaint handling.

The results of the learning experiments clearly indicate that model overfitting has already occurred at lower epoch number increases, which is a particularly critical aspect for systems processing bank data where generalization ability is key (Figure 3.). The experimental phase was run with a batch size of 16 and a training cycle of 8 epochs. In the initial tests, I first increased the number of epochs, with a targeted focus on the impact of each increase on validation performance and on the onset of over-learning.

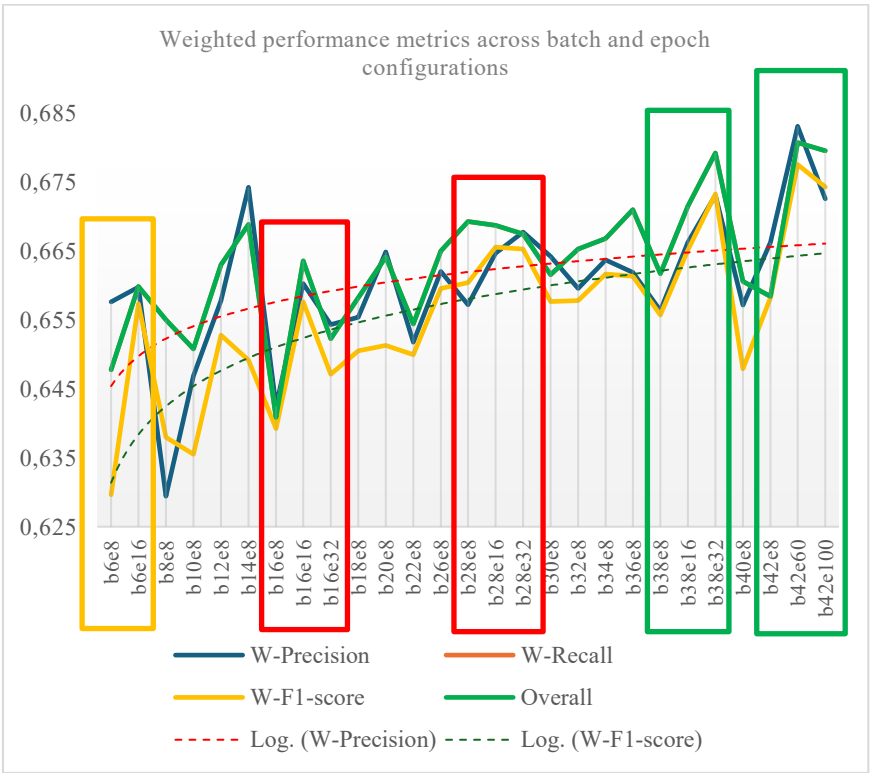

Figure 3 – Overfitting and batch sizes

Although the figure is labelled for consistency with earlier internal evaluations, the primary metrics visualized are Weighted Precision and Weighted F1-score rather than overall accuracy, in line with the risk-sensitive nature of banking complaint classification. After detecting the limits of overfitting, the model was trained along a two-way batch size optimization. First, by gradually decreasing the batch size, and then by gradually increasing the batch size, I tested the extent to which the predictive performance of the model could be maintained, with a particular focus on the stability of the weighted metrics (W-Precision, W-F1). The training was performed under the specificity of the bank dataset, which placed extreme memory demands on the computational infrastructure.

In order to explore the memory constraints of the training environment, I examined the Out of Memory (OOM) tolerance on a workstation specifically optimized for artificial intelligence workloads and equipped with a GPU featuring 24 GB of VRAM. For transformer-based models, the GPU memory required per training iteration can be well approximated as a function of the batch size (B), the maximum sequence length ($L_{\max}$), and the number of model parameters ($|\ \theta\ |$), according to the following relation:

$$M_{\text{total}} \approx M_{\text{param}} + B \cdot M_{\text{activ}}(L_{\max}),$$

where M_{param} denotes the memory required to store the model parameters, and M_{activ} represents the memory demand associated with activations and gradient values per sample. With the maximum sequence length fixed ($L_{\max} = 512$), the batch size increases memory consumption linearly, thereby directly determining the OOM boundary.

Based on empirical measurements, it was determined that under the available 24 GB of VRAM, the maximum batch size that could be handled during training was $B_{\max} = 42$. Beyond this value, the GPU memory capacity proved insufficient to execute the full forward and backward propagation steps. Accordingly, the batch size used during training had to be constrained by this upper limit.

The effect of batch size scaling on learning dynamics and performance metrics was investigated using a systematic experimental design. For both reduced and increased batch sizes, training was performed using four different epoch counts, which enabled a detailed analysis of the relationship between convergence behavior and generalization performance.

The results clearly indicated that smaller batch sizes – particularly when processing formally structured text corpora characteristic of the banking domain – led to an increased tendency toward overfitting. This phenomenon manifested in the fact that increasing the number of epochs did not result in meaningful performance improvements, while stagnation or degradation of the validation loss was observed. Mathematically, this behavior can be described by the divergence between the training and validation losses:

$$\lim_{t \to T} (\mathcal{L}_{\text{train}}(t) - \mathcal{L}_{\text{val}}(t)) \uparrow,$$

where t denotes the epoch index.

Consequently, for configurations with batch sizes $B < 38$, no further increase in the number of epochs was applied, as the deterioration in generalization performance could not be compensated by longer training cycles. This observation is consistent with findings reported in the literature, which indicates that for narrow-domain, highly structured text corpora, the combination of small batch sizes and extended training cycles entails an increased risk of overfitting, while yielding only marginal improvements in predictive performance.

Based on the measurements and model validation results obtained, it can be concluded that for training with larger batch sizes (especially at values 38 and 42), significant improvements in classification performance indicators can be obtained with increasing the epoch number. This trend was particularly evident for Weighted Precision and Weighted F1-score, which are of primary importance in the automatic processing of textual reports, which are formally structured and require high precision, typical of the financial sector.

In the subsequent analyses, I placed particular emphasis on the quantitative examination of the Cost/Income (C/I) ratio, which contrasts the computational and energy requirements associated with model training against the achievable gains in predictive performance. The cost side was characterized based on empirical measurements: under the configuration with batch size $B = 32$ and epoch count $e = 8$, with a maximum sequence length of $L_{\max} = 512$ tokens, the total training runtime was approximately 4.6 hours, while the average GPU power consumption was around 307 W. This corresponds to a total GPU-side energy consumption of approximately 1.41 kWh, resulting in an additional cost of roughly 0.176 kWh per epoch.

The income side was characterized by the evolution of weighted performance metrics, primarily weighted precision and the weighted F1-score. A comparison across different training configurations revealed that while computational and energy demands increase approximately linearly with the addition of epochs, the achievable performance gains follow a substantially slower, saturating trajectory. This relationship is illustrated by a logarithmic trend line fitted to the measured data, which approximates the relationship between performance gain and computational effort as

$$\Delta\text{Performance}(C) \approx a \cdot \ln(C) + b,$$

where C denotes the cumulative computational cost (in this case, GPU-side energy consumption measured in kWh), and a and b are constants determined through empirical fitting. The logarithmic trend line was fitted to the measurement data using a spreadsheet environment (Microsoft Excel), employing the least squares method to estimate the fitting parameters.

The flattening of the trend line clearly indicates the phenomenon of diminishing returns: beyond a certain configuration, the additional energy consumption of approximately 0.176 kWh per epoch yields only marginal improvements in the weighted performance metrics. This quantitatively demonstrates that the C/I ratio deteriorates beyond a specific training regime, meaning that further computational investment results in disproportionately small predictive gains. Accordingly, when selecting the model configuration, optimizing the empirically observed cost-benefit trade-off is more justified than pursuing maximum attainable performance.

It should be noted that recall-oriented performance is not ignored in the proposed system, complaints that are not classified with sufficient confidence are explicitly routed to manual processing, ensuring full coverage of customer submissions without compromising regulatory reliability.

3.1 Scientific aspect and relevance of the results

3.1.1 Development of a bank-specific NLP model

One of the most significant innovations of the research was the development of a language model specifically designed for the financial context, capable of automatically classifying unstructured banking customer messages – typically complaints and other transactional communications – with high accuracy. It will enable a significant speed-up and partial automation of customer service processes, while reducing the risk of human error.

3.1.2 Use of a large volume, real-world banking dataset

The model has been trained and validated on a unique Hungarian-language database of over 129,000 records of real customer complaints. The use of such a large and authoritative dataset has provided the opportunity to create a highly specialized NLP model for financial purposes, which, to our knowledge, has not been published in national or international literature. This represents a significant scientific added value, especially in the field of language-specific (Hungarian) financial AI solutions.

3.1.3 Optimized learning strategy and scalability

The detailed hyperparameter tuning carried out in this research, in particular the joint optimization of batch size and epoch number, showed that with larger batch sizes, model convergence occurs later, but the risk of over-learning can be reduced. This experience can guide the design of future learning strategies for financial language models, in particular with respect to cost/income ratio and sustainability. One of the main practical implications of this research is, therefore, how to optimize the trade-off between resource requirements and predictive performance in a scientifically sound way.

3.1.4 Targeted classification of customer complaints and transaction messages

The developed model is not only suitable for general text analysis tasks but also performs very effectively in structuring customer communication in Hungarian banking. This is of particular importance because misclassified complaint messages in the financial sector can have a direct impact on internal processes, regulatory compliance (e.g. MNB statistics) and customer satisfaction. The model is, therefore, not only a technological innovation but also a valuable business tool to improve risk management and customer experience.

3.1.5 Scientific and industrial impact

The results of this research will contribute to the development of the applied artificial intelligence and natural language processing disciplines, in particular towards specialized, domain-specific (banking) applications. It may also provide practical guidance for financial service providers wishing to implement or further develop AI-based customer communication systems. The paper also shows that Hungarian-language NLP applications are able to provide competitive performance in international practice, given appropriate data quality and learning strategy.

4 Discussion

The experience gained in the model development and fine-tuning presented in this research sheds new light on the design and applicability of natural language processing systems tailored to the financial sector in several respects. Empirical data obtained from DistilBERT-based model training and optimization support the hypothesis that the efficiency and predictive performance of language models depend significantly on the precise calibration of training parameters (mainly batch and epoch sizes), especially for domain-specific (e.g., banking) texts in Hungarian.

The phenomenon of overlearning in training at lower batch sizes shows that the "smaller batch - better generalization" paradigm of traditional machine learning does not necessarily hold for finely segmented texts in financial contexts. The model performance showed stability towards larger batch sizes, suggesting that in this application domain, larger sample size helps detect representative patterns and filter out noise. However, this approach requires increased computational capacity, which implies a resource-intensive learning strategy and raises issues of scalability and sustainability.

The predictive accuracy, in particular the weighted precision and the weighted F1-score, shows that the model performed exceptionally well in the automatic classification of unstructured complaint texts, confirming its practical relevance in banking customer service. The secondary role of the recall scores in the study does not diminish their importance, but the specificities of the financial sector, in particular the requirements for data quality, data traceability and regulatory compliance, justify considering the accuracy of the classification as a primary indicator.

The logarithmic trend analysis used in the learning process clearly identifies the phenomenon of "diminishing returns", i.e. an increase in additional computational resources does not lead to a proportional increase in accuracy. This is a particularly important conclusion for economic sustainability since the analysis of the Cost/Income (C/I) ratio suggests that the optimal model configuration is not

necessarily one that aims at achieving absolute maximum accuracy, but rather an ideal balance between predictive performance and resource requirements.

Of particular note is that the model has been trained and validated exclusively on real client data in Hungarian, which is rare in current NLP literature and contributes to the linguistic and cultural diversity of international research. This research not only extends the applicability of NLP models in technological terms, but also contributes to the regional financial digitization efforts. From a regulatory and operational perspective, model interpretability represents an important complementary aspect of automated complaint classification systems. While the present study focuses primarily on predictive performance and efficiency, future work will investigate the integration of post-hoc explainability techniques, such as attention analysis or SHAP-based token attribution, to support auditability, internal validation, and regulatory transparency in banking environments.

Finally, while the results are promising, the research has natural limitations: the training was done on a closed data set and the model in its current form is not tested in a real-time, online environment. Future research should extend these dimensions, with a particular focus on issues of generalizability, adaptivity and multilingualism.

Conclusions

The aim of the present research was to create and fine-tune a natural language processing model for the Hungarian banking context, capable of automatically classifying unstructured customer communications, in particular complaint reports. The DistilBERT architecture used in the development of the model allowed to achieve an optimal balance between accuracy, computational efficiency and cost-effectiveness. By using large volumes of real-world banking customer data, the model is not only theoretically novel, but also of practical relevance for the digitalization of financial services and the improvement of customer satisfaction.

One of the main contributions of this research is the optimization of the learning strategy, with a special focus on the detailed study of the behavior of epoch-batch parameter pairs in a banking NLP task. The results show that with a larger batch size, the risk of over-learning is lower and more epochs are required for convergence - a finding that may open up new perspectives for streamlining the training process of language models with similar goals.

Furthermore, the research has also focused on sustainability considerations. By analyzing the relationship between cost-effectiveness and predictive performance, I have shown that further improvements in predictive accuracy are associated with exponentially increasing computational and financial overheads, so that the phenomenon of "diminishing returns" is well understood in the development of financial NLP models.

References

[1] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” Science, vol. 349, no. 6245, pp. 261–266, 2015.

[2] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson, 2023.

[3] A. J. Tsai et al., “Financial text mining: A review,” Journal of Banking & Finance, vol. 127, 2021.

[4] S. Loughran and B. McDonald, “Textual analysis in accounting and finance: A survey,” Journal of Accounting Research, vol. 54, no. 4, 2016.

[5] M. A. Haque et al., “Automated complaint classification using machine learning,” IEEE Access, vol. 7, pp. 136800–136812, 2019.

[6] T. Kiss and P. Vincze, “Challenges in Hungarian natural language processing,” Computational Linguistics, vol. 45, no. 2, 2019.

[7] C. D. Manning et al., “Introduction to information retrieval,” Cambridge University Press, 2008.

[8] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” NAACL, 2019.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” NeurIPS, 2019.

[10] M. Huang et al., “Financial BERT: A pre-trained language model for financial NLP tasks,” EMNLP, 2021.

[11] Magyar Nemzeti Bank, “57/2023. (XI. 24.) rendelet a panaszkezelési szabályokról,” 2023.

[12] A. Smith and B. Johnson, “Efficient fine-tuning of transformer models for domain-specific tasks,” ICML, 2022.

[13] R. Wilson et al., “Optimizing batch size and epochs for deep learning models,” Journal of Machine Learning Research, vol. 22, 2021.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?” Explaining the Predictions of Any Classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[15] A. Cambria and B. White, “Jumping NLP Curves: A Review of Natural Language Processing Research,” IEEE Comput. Intell. Mag., vol. 9, no. 2, pp. 48–57, 2014.

[16] L. Vincze et al., “Szintaktikai elemzés magyar nyelvű szövegeken: kihívások és lehetőségek,” Infokommunikáció és Jog, vol. 21, no. 3, pp. 34–42, 2022.

[17] D. Németh, G. Recski, and K. Kornai, “Training BERT Models for Hungarian: The Case of HuBERT,” in Proc. 12th Language Resources and Evaluation Conf. (LREC), 2020.

[18] T. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems (NeurIPS), 2020, vol. 33, pp. 1877–1901.

[19] P. Felföldi et al., “PULI-GPT: Hungarian Large Language Model Development,” arXiv preprint arXiv:2403.09350, 2024.

[20] I. Beltagy, M. Peters, and A. Cohan, “Longformer: The Long-Document Transformer,” arXiv preprint arXiv:2004.05150, 2020.

[21] E. Strubell, A. Ganesh, and A. McCallum, “Energy and Policy Considerations for Deep Learning in NLP,” in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 3645–3650.

[22] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations (ICLR)*, 2015.