Degree Project in Technology

Second cycle, 15 ETCS

Deep Learning Techniques for Bank Transaction Categorization

A collaborative study with Kreditz

FILIP ORESTAV & KOLUMBUS LINDH

Deep Learning Techniques for Bank Transaction Categorization

A collaborative study with Kreditz

FILIP ORESTAV & KOLUMBUS LINDH

Svensk titel:
Djupinlärningstekniker för Kategorisering av Banktransaktioner - En Samarbetsstudie med Kreditz

Sammanfattning:
Djupinlärning (DL) har stor potential för att förbättra noggrannheten i kategorisering av bank-transaktioner. Denna studie utvärderar effektiviteten av djupinlärningsmodeller för att identifiera latenta mönster inom banktransaktioner, med målet att förbättra prestandan hos Kreditz AB kategoriseringsmotor. Vår metod innefattar användningen av avancerade DL-tekniker och den omfattande transaktionsdatamängd som tillhandahålls av Kreditz. Vi använder teckennivå-inbäddningar för att bearbeta textdata, vilket möjliggör fångandet av komplexa relationer med minimala textinmatningar, vilket är idealiskt för korta banktransaktionsbeskrivningar.

Våra resultat visar att faltningsnätverk (CNN) levererar topprestanda vad gäller F1-poäng samt tränings- och inferenshastigheter. Även om Lång Kortsiktigt Minne-nätverk (LSTM) matchar F1-poängen för CNN, gör de det i en långsammare takt. Multi-Layer Perceptrons (MLP), även om de är snabbast, uppnår inte jämförbara F1-poäng.

Den kombinerade höga hastigheten och noggrannheten hos dessa modeller möjliggör värdefulla funktioner som nästan realtidskategorisering av transaktioner med hög precision. Denna förmåga har betydande implikationer, inklusive potentialen att mer effektivt bedöma kreditpoäng, förebygga penningtvätt och hantera spelberoende.

Deep Learning Techniques for Bank Transaction Categorization

A Collaborative Study with Kreditz

Filip Orestav & Kolumbus Lindh

KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science, EECS

June 3, 2024

***Abstract*—Deep learning (DL) holds significant promise for enhancing the accuracy of bank transaction categorization. This study evaluates the effectiveness of deep learning models in identifying latent patterns within bank transactions, with the aim of improving the performance of Kreditz AB's categorization engine. Our approach involves the use of advanced DL techniques and the extensive transaction dataset provided by Kreditz.**

We employ character-level embeddings to process textual data, enabling the capture of complex relationships with minimal text input, which is ideal for concise bank transaction descriptions. Our findings reveal that Convolutional Neural Networks (CNNs) deliver state-of-the-art performance in terms of F1-score, as well as training and inference speeds. While Long Short-Term Memory networks (LSTMs) match the F1-score of CNNs, they do so at a slower pace. Multi-Layer Perceptrons (MLPs), although the fastest, fail to achieve comparable F1-scores.

The combined high speed and accuracy of these models facilitate valuable features such as near real-time transaction categorization with high precision. This capability has significant implications, including the potential to more effectively assess credit scores, prevent money laundering and address gambling addiction.

***Index Terms*—Deep Learning, Multilayer Perceptron, Convolutional Neural Network, Long Short Term Memory, Bank Transaction Categorization, Machine Learning, Credit Scoring, Anti-Money Laundering**

I. Introduction

Machine learning (ML) is revolutionizing industries and reshaping our professional landscapes. Its role varies—sometimes a tool, other times a replacement—but its permanence is undeniable. A distinctive advantage of ML is its scalability with data: the more data it receives, the more effectively it can operate. In our data-rich environment, this characteristic is particularly potent.

Increasingly, companies are leveraging various ML models to enhance their business offerings, with some even centering their entire product around these technologies. Kreditz AB exemplifies this trend, harnessing consent-based data coupled with AI technologies to empower their clients. By providing meticulously categorized and structured bank transaction data, Kreditz enables informed, intelligent decision-making processes for businesses relying on financial insights.

A. Motivation and challenge

Founded in 2018, Kreditz has rapidly expanded its influence in the financial technology sector. With over 40 clients and more than 500 million transactions analyzed [1], the company leverages bank transaction data to enhance business-to-consumer interactions, facilitating more informed decision-making processes.

The application of transaction analysis is multifaceted. For banks, it serves as a powerful tool for verifying income, spending behaviors, as well as assessing assets and debts, thereby streamlining the lending process. Businesses benefit from insights into payment flows and tax obligations, and can monitor financial behaviors more accurately. Additionally, this analysis is crucial in identifying gambling addictions by detecting spending anomalies.

Despite the successes, Kreditz seeks to explore how deep learning models can augment their existing categorization engine. Historically, traditional machine learning models, such as XGBoost, have outperformed deep learning in this domain. However, there is a belief that deep learning could unlock new potentials, particularly in specific transaction categories, thereby enriching the insights gained from the data.

Currently, Kreditz's categorization engine achieves an accuracy rate of approximately 95% in transaction classification. Although this is commendable, there is a continuous drive to surpass this benchmark. Enhancing accuracy, even marginally, is pivotal—not only for maintaining competitiveness but also for attracting new clients in a market where rivals boast similar capabilities. This research aims to support Kreditz in achieving these goals by integrating cutting-edge deep learning techniques to refine and perfect transaction categorization.

B. Research objective

The primary aim of this study is to explore how deep learning technologies can enhance the accuracy of transaction categorization at Kreditz. Eight categories on the incoming side (inflowing transactions) are used, but the focus lies on the seven most important categories, trying to improve classification accuracy and precision, as well as identifying new patterns defining those categories.

Existing technologies in the field often rely on naive approaches, such as keyword matching or manual feature extraction. Deep learning models, however, offer the potential to capture more complex relationships with less human intervention.

This project aims to develop and train deep learning models capable of facilitating new market entries by training on a relatively small dataset, similar to what Kreditz might have when entering new markets. The performance will determine its effectiveness as a preliminary tool before developing a more complete engine. Additionally, the model is expected to act as a fallback layer for data points that the current engine cannot classify. We will assess various deep learning models to identify the most effective approach, evaluating the potential benefits and implications for Kreditz's value creation and proposition.

Focusing on a subset of transactions, this research serves as a proof of concept to demonstrate how companies like Kreditz can effectively utilize this technology.

C. Contribution to society

This research not only aims to improve bank transaction categorization but also aligns with several of the Sustainable Development Goals, seen in Table I.

Sustainable Development Goal	**How our research supports the goal**
9: Industry, innovation and infrastructure	Facilitating the access to financial services and markets.
10: Reduced inequalities	Promoting economic inclusion.
12: Responsible consumption and production	Facilitating for companies to apply sustainable methods and sustainability reporting.
16: Peace, justice and strong institutions	Combating organized crime and illegal financial and arms flows.

TABLE I
SUPPORT OF THE SUSTAINABLE DEVELOPMENT GOALS BY OUR RESEARCH

II. PROBLEM DEFINITION

The problem is defined from two angles, one from the computer science perspective and one from the industrial engineering perspective. Both disciplines are covered simultaneously throughout the report, with an extra focus on the industrial engineering perspective in the discussion.

a) Computer science question: **To what extent can the application of DL models enhance the bank transaction categorization process and find patterns not found by ML models?**

b) Industrial engineering question: **What are the benefits of using DL in this field, and especially for Kreditz?**

A. Technical feasibility

Data quality and availability: The proprietary access to Kreditz's dataset mitigates a common challenge in DL research: the availability of labeled datasets. This unique access allows for the training and validation of DL models on a comprehensive and relevant dataset, setting this research apart from studies constrained by public or less specific datasets. It should however be noted that the quality of the dataset is somewhat reduced, since GDPR-compliance prevents us from using raw data.

Model complexity vs. interpretability: Advanced DL models, while potentially offering enhanced accuracy, often suffer from a lack of interpretability and being seen as a "black box". This poses a significant challenge in sectors where understanding decision-making processes is critical for compliance and trust.

Handling imbalanced data: The nature of financial transactions, where certain transactions are far more common than others, presents a challenge in training DL models as the models easily become biased towards the majority classes. In this research, we use class weights during the training process to penalize mistakes in the majority classes more than majority classes, leading to more balanced models.

III. THEORY & PRIOR STUDIES

A. Prior studies and empirical findings

The integration of DL technologies, including LSTM, CNN, GRU, and Transformer models, has significantly advanced bank transaction categorization. Empirical studies show these DL models outperform traditional machine learning methods in handling complex datasets [4], [5], [6], [7]. Recent advancements highlight DL's crucial role in enhancing data-driven decision-making within financial services. Studies by Sawan Rai (2021)[1] and Vasista Reddy (2023)[2] demonstrate the superiority of LSTM, GRU, and CNN models over traditional models like Logistic Regression, Support Vector Machine, and Naive Bayes, especially with large datasets. These findings underscore DL's effectiveness in extracting intricate patterns from transaction data and its scalability in processing extensive financial datasets [8], [5], [7].

CNNs have evolved from being peripheral to financial applications to excelling in transaction categorization, fraud detection, and customer behavior analysis due to their hierarchical feature extraction capabilities. This evolution is supported by studies that combine CNNs with attention mechanisms for short text classification, outperforming traditional techniques [9], [10].

Empirical evidence shows LSTM and GRU models are essential for predictive modeling in finance due to their ability to model long-term dependencies. Recent advancements in computational power and data availability have increased their application, marking a shift from their initial cautious adoption [8], [5].

B. Models to be explored

This research explores several models: Multilayer Perceptrons (MLPs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) Networks, and Convolutional Neural Networks (CNNs).

1) Multilayer Perceptrons (MLPs): MLPs are neural networks where each node is fully connected to nodes in the next layer. They consist of at least three layers: an input layer, one or more hidden layers, and an output layer.

[1]Rai, S. (2021). Empirical Study of Traditional vs. Deep Learning Classification Models. https://www.analyticsvidhya.com/blog/2021/05/emperical-study-of-traditional-vs-deep-learning-classification-models/

[2]Reddy, V. (2023). Deep Learning vs. Machine Learning with Text Classification. https://vasista.medium.com/deep-learning-vs-machine-learning-with-text-classification-162ea20a7924

2) Recurrent Neural Networks (RNNs): RNNs are designed to process sequential data, such as transaction histories. They remember past inputs using their internal state, crucial for tasks requiring knowledge of previous data points to predict future ones [11].

RNNs face challenges with long sequences due to the vanishing gradient problem, where the gradient of the loss function decreases exponentially with the length of the input sequence, making it hard to learn long-range dependencies.

a) LSTM Networks: LSTM networks, a type of RNN, address the vanishing gradient problem with a specialized architecture that includes forget, input, and output gates to control the flow of information. This makes LSTMs better at managing longer sequences [12]. Because LSTM is effective at analyzing sequential data, this can be particularly helpful when understanding the context of a certain combination of characters.

b) BiLSTM Networks: A variant of LSTM, called Bidirectional LSTM (BiLSTM), is also explored. BiLSTM has two LSTM-layers and processes input both forward and backward, combining outputs from both directions. This ability can give the model an even better understanding and context of the sequential data in the transaction descriptions.

3) Convolutional Neural Networks (CNNs): CNNs are designed for image analysis, processing data in smaller regions with small learnable filters called kernels. These kernels slide across the image, performing element-wise multiplication with the input data to identify local patterns, allowing CNNs to perform dynamic image categorization.

Beyond image analysis, CNNs have shown success in various tasks due to their ability to extract patterns from data. In financial transaction categorization, CNNs have demonstrated promising results, potentially due to their effectiveness in handling unstructured data with multicollinearity [13].

C. Financial benefits of improved categorization

Accurate transaction categorization offers significant financial benefits, reducing costs in several areas:

1) Reduced fraudulent transactions: Enhanced transaction categorization accuracy helps minimize financial losses from fraudulent activities. Deep learning techniques for fraud detection, as explored by Ngai et al. (2011), are effective in identifying anomalies in transaction patterns [14]. This allows for faster detection and intervention, potentially saving financial institutions billions annually, as estimated by The Nilson Report (2023) [15].

2) Enhanced financial management: Accurate transaction categorization gives individuals deeper insights into their spending patterns, facilitating informed budgeting decisions and potential cost savings. Research by Agarwal et al. (2019) indicates that transaction categorization tools can significantly reduce unnecessary spending among young adults [16]. By providing users with a clear understanding of their expenses, these tools encourage more responsible spending habits.

D. Potential societal impact

Beyond financial benefits, improved transaction categorization has broader societal implications, contributing to economic stability and social welfare:

1) Financial inclusion: Accurate transaction categorization can help expand access to financial services for underserved populations by enabling more accurate risk assessments. Research by Demirgüç-Kunt et al. (2018) shows that alternative data sources, such as transaction data, can help overcome traditional barriers to credit access, particularly in developing economies [17].

2) Combatting money laundering: Precise categorization of financial transactions is crucial for detecting and preventing money laundering activities, ensuring the integrity and stability of the global financial system. Studies like Cheng et al. (2017) highlight the importance of accurate transaction categorization in identifying suspicious activities and enhancing regulatory compliance [18].

IV. Dataset

Our dataset consisted of 200 000 bank transactions from Swedish banks which were labeled according to Kreditz's current categorization engine. This meant that we could not exceed the engine in terms of accuracy, but mainly in terms of speed. All transactions had positive amounts, and no information about the client was present in the transactions. Each transaction adhered the format shown in Table II.

ID	transaction_date	transaction_amount
The ID of the transaction	YYYY-MM-DD	Swedish krona

transaction_description	transaction_class
Short description, e.g. 'avanza', 'swish', etc.	Target label

TABLE II
FORMAT OF OUR DATA

The transaction classes were distributed and mapped as shown in Table III.

Label	Actual class	Distribution (%)
class1	Child benefit	0.930
class2	Bonus (casino)	0.003
class3	Salary	4.111
class4	Pension	0.790
class5	Maintenance support	0.138
class6	Dividend payout	0.010
class7	Transfer between banks	13.046
others	Internal bank-transfers, tax, insurance, etc.	80.974

TABLE III
DISTRIBUTION AND MAPPING OF CLASSES/LABELS IN THE DATASET

V. Methodology

A. Pre-processing the data

The first step of the implementation involved data cleaning and preparation. Kreditz's data was highly pre-processed to comply with GDPR regulations. Nonetheless, it was necessary to drop transactions with NaN values. Additionally, some transaction descriptions were identical but found in different classes, primarily in class 7 and the "others" class. These transactions were removed to train the models on clean data and obtain the most representative evaluation.

B. Embeddings

Due to the high level of pre-processing, the text data was often incomplete and misspelled. Therefore, we opted for character-level embeddings instead of the more common word-level embeddings. Character-level embeddings enable the model to identify sequences of tokens or words that are similar, even if they are misspelled or missing characters.

The textual data was represented in a vector format with a maximum length of $C = 34$, corresponding to the largest sequence of characters in the dataset. Each character token was encoded and mapped to an integer, where each integer served as an index mapping the token from the vocabulary.

Additionally, numerical and text data were converted into vector embeddings. The transaction amounts were scaled using min-max scaling. We chose to one-hot encode the day of the week, as transaction patterns often vary by day, such as between weekdays and weekends. The remaining numerical data was encoded ordinally to preserve their inherent ordering.

Each bank transaction was in the end represented in the following vector format:

$$[X_{C1}, X_{C2}, ..., X_{C34}, X_{\text{Amount}},\\ X_{\text{Day of month}}, X_{\text{Month number}}, OHE_{\text{Day of week}}]$$

C. Training procedure

The data was split into 80% training and 20% testing sets. The training data was further divided using k-fold cross-validation with five folds, where each fold consisted of 80% training data and 20% validation data. Thus, the training data consisted of 120,000 samples, the validation data 24,000 samples, and the test data 20,000 samples. Each set maintained the same class distribution.

To address the imbalanced dataset, class weights were used during the training process. This approach penalized mistakes in minority classes more than in majority classes.

Each fold was trained for a maximum of 30 epochs. Early stopping with a patience of 10 epochs monitored the validation loss, and the best weights were restored if no improvement was observed.

D. Training parameters

All models were trained using a batch size of 256 and the Adam optimizer algorithm with a learning rate $\alpha = 0.001$.

Since this was a multiclass classification task, the categorical cross-entropy loss function was chosen:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{8} y_i \log(\hat{y}_i) \tag{1}$$

Here, $\mathbf{y}$ represents the true distribution of classes, where $y \in \{1, 2, 3, 4, 5, 6, 7, 8\}$. $\hat{\mathbf{y}}$ represents the predicted distribution, which the model aims to match as closely as possible to $\mathbf{y}$. The loss is summed over all classes, with y_i and $\hat{y}_i$ representing the i-th elements of $\mathbf{y}$ and $\hat{\mathbf{y}}$, respectively.

E. Models

TABLE IV
NUMBER OF PARAMETERS FOR EACH MODEL

Model	Number of parameters
MLP	974
CNN	1182
LSTM	798
BiLSTM	1438

We started with minimalist architectures and increased complexity only when necessary to address underfitting. The depth and width of our networks were tuned to find the optimal balance between underfitting and overfitting.

Our strategy was to use the minimum depth and width required to maintain performance while reducing training and inference times. This approach was effective, but we acknowledge that more complex datasets may necessitate additional layers to capture nuanced patterns.

All our models utilized a bifurcated approach, with separate branches for textual and numerical data. The textual data was processed through an embedding layer, while the numerical data passed through a dense layer. For all models tested, both the embedding and the numerical dense layers had a fixed output length of 3.

These branches were then concatenated before the final dense layer, which used the softmax activation function (see equation (2)) to generate a probability distribution.

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{8} e^{z_j}} \tag{2}$$

where z_i represents the i-th element of the input vector z to the softmax function, and 8 is the total number of classes. This approach ensures that the final output is a probability distribution over the possible transaction categories.

1) Multilayer Perceptron (MLP): The MLP architecture is outlined in Table V and visually represented in Figure 1. It is designed to be minimalistic, using as few layers as possible.

The shallow design of the MLP aims to prevent overfitting, promoting good generalization and fast training. However, this simplicity may limit its ability to capture complex relationships in more intricate datasets.

Layer	Input Units	Output Units
Text Input	[34 × 1]	[34 × 1]
Embedding	[34 × 1]	[34 × 3]
Flatten	[34 × 3]	[102 × 1]
Numerical Input	[10 × 1]	[10 × 1]
Dense (Numerical)	[10 × 1]	[3 × 1]
Concatenate	[102 × 1], [3 × 1]	[105 × 1]
Dense (Output)	[105 × 1]	[8 × 1]

TABLE V
INPUT AND OUTPUT UNITS FOR EACH LAYER IN THE MLP MODEL

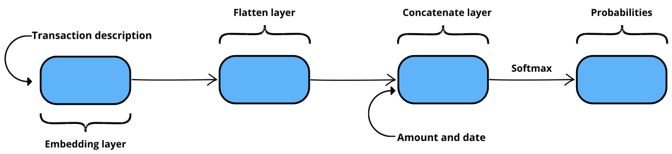

Fig. 1. MLP architecture.

2) Convolutional Neural Network (CNN): The CNN architecture offered more parameters for adjustment compared to the MLP, owing to its increased complexity. Details of the layers and their dimensions can be found in Table VI, with a visual overview provided in Figure 2.

Layer	Input Units	Output Units
Text Input	[34 × 1]	[34 × 1]
Embedding	[34 × 1]	[34 × 3]
Conv1D	[34 × 3]	[34 × 16]
MaxPooling1D	[34 × 16]	[6 × 16]
Flatten	[6 × 16]	[96 × 1]
Numerical Input	[10 × 1]	[10 × 1]
Dense (Numerical)	[10 × 1]	[3 × 1]
Concatenate	[96 × 1], [3 × 1]	[99 × 1]
Dense (Output)	[99 × 1]	[8 × 1]

TABLE VI
INPUT AND OUTPUT UNITS FOR EACH LAYER IN THE CNN

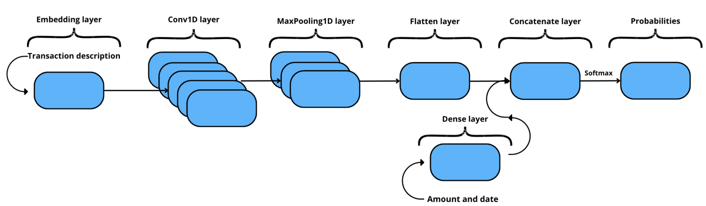

Fig. 2. CNN architecture.

3) Long Short-Term Memory (LSTM): The architecture and parameters of the LSTM were kept minimal, similar to the other models. The LSTM layer had 10 units, chosen to evaluate its ability to store more information from the input sequence, potentially capturing more complex patterns and long-term dependencies. Details of all layers and their dimensions for the LSTM model are provided in Table VII, with a graphical representation in Figure 3.

Layer	Input Units	Output Units
Text Input	[34 × 1]	[34 × 1]
Embedding	[34 × 1]	[34 × 3]
LSTM	[34 × 3]	[10 × 1]
Numerical Input	[10 × 1]	[10 × 1]
Dense (Numerical)	[10 × 1]	[3 × 1]
Concatenate	[10 × 1], [3 × 1]	[13 × 1]
Dense (Output)	[13 × 1]	[8 × 1]

TABLE VII
INPUT AND OUTPUT UNITS FOR EACH LAYER IN THE LSTM

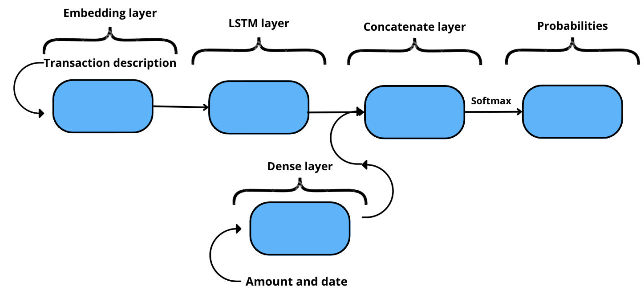

Fig. 3. LSTM architecture.

4) Bidirectional Long Short-Term Memory (BiLSTM): The BiLSTM model shared many similarities with the LSTM model. However, it had 20 units in the LSTM layer (due to two LSTM layers processing input, both left to right and right to left). Details of the architecture, including all layers and their dimensions, are provided in Table VIII, with a graphical representation in Figure 4.

Layer	Input Units	Output Units
Text Input	[34 × 1]	[34 × 1]
Embedding	[34 × 1]	[34 × 3]
Bidirectional LSTM	[34 × 3]	[20 × 1]
Numerical Input	[10 × 1]	[10 × 1]
Dense (Numerical)	[10 × 1]	[3 × 1]
Concatenate	[20 × 1], [3 × 1]	[23 × 1]
Dense (Output)	[23 × 1]	[8 × 1]

TABLE VIII
INPUT AND OUTPUT UNITS FOR EACH LAYER IN THE BILSTM

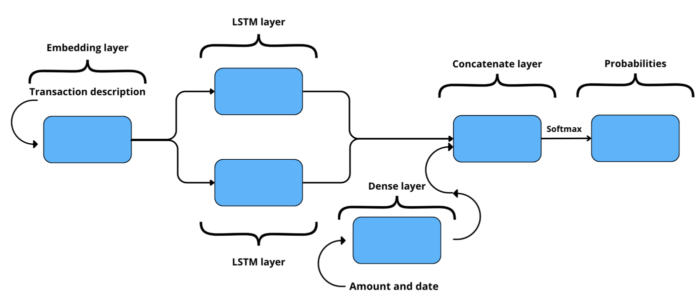

Fig. 4. BiLSTM architecture.

F. Evaluation

Given the imbalanced nature of our dataset, we used the macro-averaged F1-score to evaluate the models' performances. This metric calculates the average F1-score across all classes, without being influenced by the support of each class. Therefore, each class is equally weighted in the evaluation.

The macro-averaged F1-score is computed as follows:

$$F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{3}$$

where N is the number of classes, and $F1_i$ is the F1-score for the i-th class, which is computed as:

$$F1_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (4)$$

G. Hardware

Training was conducted on M-series Apple Silicon chips (M1 and M2). Most models were trained in under 10 minutes, enabling experimentation with various layer and hyperparameter configurations. However, we did not perform exhaustive grid searches due to resource constraints that would have increased training time beyond our available resources.

VI. RESULTS

A. Macro averaged F1-scores and speed

TABLE IX
MACRO AVERAGE SCORES AND RELATIVE TIMES FOR OUR TESTED MODELS

Model	Macro Averaged F1-Score	Training Time (relative)	Prediction Time (relative)
MLP	0.93	**1**	**1**
CNN	**0.97**	2.43	1.08
LSTM	**0.97**	7.52	2.15
BiLSTM	0.78	14.37	3.15

The four models tested—MLP, CNN, LSTM, and BiLSTM—demonstrated strong capabilities in predicting transactions. As shown in Table IX, the CNN and LSTM performed the best among the models in terms of macro-averaged F1-score.

The speed of the models exhibited remarkable differences in both training and inference. The MLP exhibited superior performance in both training and inference speed. Following the MLP, the CNN showed promising speed. However, the LSTM and BiLSTM models were the slowest, being 7-14 times slower than the MLP during training and 2-3 times slower during inference.

B. Confusion matrices

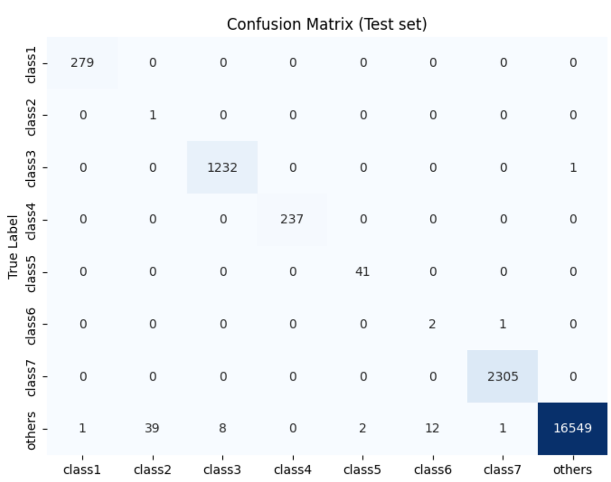

Fig. 5. Confusion matrix for BiLSTM

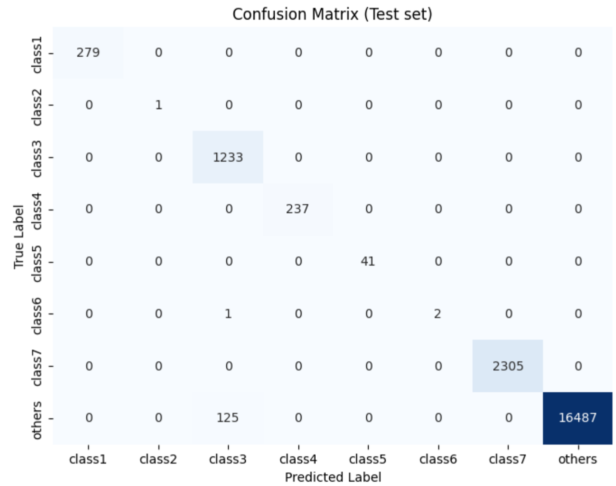

Fig. 6. Confusion matrix for CNN

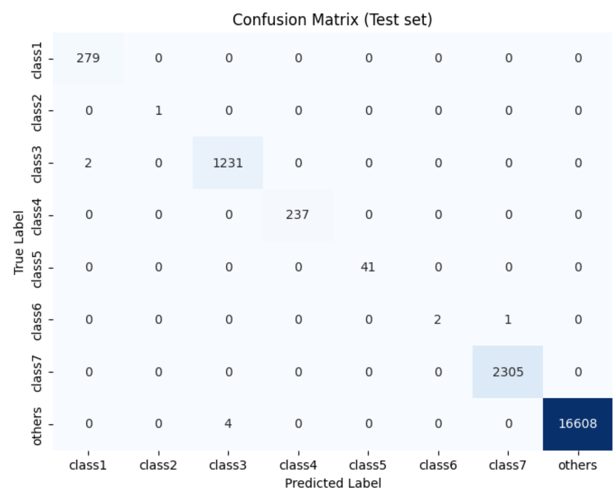

Fig. 7. Confusion matrix for LSTM

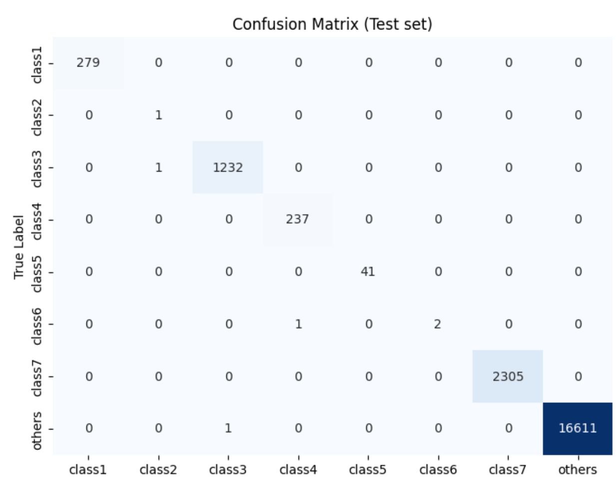

Fig. 8. Confusion matrix for MLP

VII. Discussion

A. LSTM vs BiLSTM

The LSTM proved to be superior to the BiLSTM at this task. With both having their internal memory capabilities, the LSTM has the ability to store the most important information in order to understand the context of past characters, while the BiLSTM uses context from both the past and future characters.

The LSTMs uses forget gates to determine which information to discard from a previous state. This capability allows it to store the relevant characters and leave attention at them. Since combinations of characters in a specific order can contain a lot of information, the sequential nature of LSTM provides value. For example, the word ”lön” (salary in Swedish) has the letters l, ö and n. In this case, the order is crucial. Otherwise, if the model did not have a sequential memory capability, the interpretations of that sequence of characters could have been ”öln” or ”nöl” which could have a different meaning in another context.

An examination of the confusion matrices (Figures 7 and 5) reveals that BiLSTM primarily struggles with false positives in classes 2 and 6. This behavior can be attributed to the inherent differences in how each architecture identifies patterns.

Specifically, class 2 encompasses casino bonus transactions, often containing the word ”bonus” in their descriptions. Class ”others” frequently includes transactions mentioning ”everydayplus” (a type of account credit). BiLSTM, by processing sequences in both directions simultaneously, might place undue emphasis on the subsequence ”us” present in both ”bonus” and ”everydayplus.” This can lead to misclassifications where transactions from ”others” are falsely categorized into class 2 due to the presence of ”us.”

Similarly, class 6 contains dividend payouts represented by the Swedish word ”utdelning.” This word ends with the suffix ”-ning,” which also appears in many terms from class ”others”, such as ”ersättning” (compensation) and ”inbetalning” (incoming payment). BiLSTM's ability to exploit patterns from both directions might lead it to misclassify transactions containing ”-ning” from ”others” into class 6.

In contrast, the LSTM, processing the sequence from left to right only, avoids these issues. The absence of similar subsequences preceding ”utd” (the first three characters of ”utdelning”) and ”bon” (the first three characters of ”bonus”) in the dataset allows the LSTM to make more confident predictions based on the complete sequence. This highlights a potential benefit of unidirectional processing in specific scenarios: by mitigating the influence of irrelevant subsequences, it can lead to more robust classifications.

It is important to acknowledge that the observed superiority of LSTM in this task might be dataset-specific. We hypothesize that with more diverse and complex textual data, the ability of BiLSTM to leverage information from both directions could lead to superior performance. Further research in this field, with larger and more varied bank transaction datasets, is warranted to explore this possibility.

B. The CNN

An analysis of the confusion matrix (Figure 6) reveals that the CNN achieves high F1-score for all classes except class 3, which corresponds to salary transactions. Upon closer inspection of transaction descriptions in this class, we observed vocabulary variations for salary payouts, including ”lön,” ”lon,” and ”loneutbet.” Notably, ”loneutbet” is similar to ”utbet,” a common term appearing frequently in descriptions for the transaction class ”others” (e.g., ”swish utbet ...”).

As discussed in the theoretical background section, the CNN utilizes a ”sliding window” approach. In our case, it employs filters with a kernel size of 5 that scan over the word embeddings to extract localized patterns. These patterns are subsequently captured through a convolution operation.

Following the convolutional layer, a max pooling layer is applied for downsampling the feature maps. This layer identifies the most prominent features by selecting the maximum value within each pooling window. It is possible that the sequence ”utbet” (also with a length of 5) consistently attains the maximum value within the pooling window for transactions containing this keyword (e.g. in the ”others” class). This could lead to a higher weight being assigned to the ”others” class, potentially explaining the misclassification of salary transactions. While this hypothesis requires further investigation, it aligns with the theoretical underpinnings of the CNN architecture.

C. The MLP

Despite its simplicity, our MLP model delivers impressive results, boasting a macro-averaged F1-score of 93%. This performance underscores its robust ability to accurately classify nearly all classes with high precision and recall.

An intriguing observation arises from the confusion matrix (Figure 8): the MLP occasionally misclassifies transactions belonging to class 3 (salary) as class 2 (casino bonus). Notably, transaction descriptions containing terms like ”lön” or ”lon” exhibit similarities with ”bonus” solely in the shared characters ”o” and ”n.” Due to the MLP's inability to discern the positional arrangement of characters, it relies primarily on these shared characters without considering their relative order. Consequently, instances where both ”o” and ”n” occur in transaction descriptions may lead to confusion within the model, as observed in this scenario.

D. Characters are all you need

Our analysis reveals that textual data plays a critical role in transaction classification for our models. This is likely due to the high similarity in both word usage and character sequences within most transaction classes. Manual inspection of the transactions confirmed that most classes lacked distinct information in terms of amounts and dates.

However, class 3 (salary) presents a unique case. In Sweden, salary payments often occur at the month's end, with transaction amounts typically exceeding 20,000 SEK. Additionally, these transactions consistently include keywords like ”lön” or ”lon.” This combination of factors (date, amount, and keywords) provides strong signals for the models, leading to confident predictions and minimal false negatives for class 3, as observed in all confusion matrices (Figures 5, 6, 7, 8).

While textual data appears to hold the most weight, we suspect that transactions occurring on the 25th of the month could pose challenges for the models. This might contribute to the observed false positives for class 3.

E. The impact of data

Our analyses underscore the profound influence of data volume on model performance. Consistently, across all models, we observe a clear correlation: limited data availability correlates with diminished model efficacy. Classes 2 and 6 pose particular challenges due to their sparse representation in the dataset. Notably, the scant support for class 2 within the training set makes it arduous for the model to grasp its intricacies. However, even a marginal increase in support, as exemplified by class 6, yields marked enhancements in classification metrics. This demonstrates the models' ability to swiftly adapt to limited data, as evidenced by their adept handling of class 5, despite its constrained representation.

While deep learning systems are often perceived as enigmatic "black boxes", the adverse impact of data scarcity can be logically elucidated. The diminished presence of training examples impedes the model's optimization for these underrepresented classes. This phenomenon is rooted in the mechanics of backpropagation, wherein classes with fewer instances contribute negligibly to gradient updates, leading to incomplete learning for these categories.

A consequence of employing cross-entropy loss is the unequal weighting of minority classes in weight updates. With fewer data points in a class, the summation terms for that class diminish, resulting in smaller overall gradients for weights associated with minority classes. Consequently, weight updates for minority classes proceed at a slower pace during each iteration. Given a batch size of 256, prevalent in our model executions, most batches lack positive examples of minority classes, thereby diminishing the informativeness of gradients.

F. Alignment with prior research

Our research findings align well with existing literature in several key areas, while also presenting unique insights. Zhao et al. [4] demonstrate the effectiveness of deep learning in transaction categorization, emphasizing these techniques' ability to detect intricate patterns in financial datasets. Our study corroborates these findings, as all our models, particularly CNNs and LSTMs, adeptly identified complex patterns in transaction data, reinforcing the potential of deep learning in this domain.

Bharambe et al. [9] highlight the prowess of convolutional neural networks in categorizing transactions by extracting significant features through their hierarchical processing architecture, which fall in line with our results.

Our study also confirms Lipton and Stein's exploration [6] of adaptive methods in real-time transaction classification. They suggest that more complex models, like LSTMs, excel in dynamic environments. However, we can also see that CNNs handle these complex data characteristics efficiently.

Additionally, our research sheds light on the practical applications of keeping the depth and width minimized in a neural network instead of adding dropout layers. By adjusting the size of the neural network to the specific data and the task, the most appropriate balance between under- and overfitting can be achieved.

G. Competitive advantage through speed

In the competitive landscape of bank transaction classification, differentiation based solely on accuracy becomes increasingly challenging. Many providers achieve comparable performance levels using similar methodologies. Consequently, Kreditz can gain a strategic edge by focusing on factors beyond accuracy, such as processing speed.

Cost reduction, while potentially attractive, can trigger price wars and erode profit margins. Speed, on the other hand, emerges as a critical differentiator. Our analysis reveals that MLP and CNN models offer significant speed advantages compared to LSTM and BiLSTM models (Table IX). These benefits stem from their less computationally expensive nature, lower memory requirements, and faster training times. LSTM models, designed to capture long-range dependencies, necessitate more complex computations.

The speed advantage extends to the inference stage. Unlike LSTMs, which process information sequentially, CNNs can analyze the entire input in parallel. This capability makes CNNs particularly well-suited for efficient inference on GPUs. Additionally, the simpler structure of MLPs translates to the fastest inference speeds among the models tested. Therefore, selecting the optimal model requires careful consideration beyond just accuracy; a balanced approach that prioritizes both speed and performance is crucial.

Integrating the highly accurate and exceptionally fast CNN or MLP model into Kreditz's categorization engine could unlock substantial value. This approach presents an opportunity to bolster the engine's security with minimal impact on processing speed. Notably, both MLP and CNN boast inference speeds measured in microseconds, making them remarkably lightweight. However, a trade-off exists: increased accuracy might lead to a slight increase in overall processing time. Kreditz should carefully evaluate its customers' time sensitivity and tolerance for such changes.

An alternative strategy could involve replacing the current engine entirely. This approach could be piloted on a limited set of transaction classes, followed by a comprehensive evaluation. If both accuracy and processing speed improve compared to the existing system, the need for a rudimentary keyword-matching technique could be eliminated, leading to significant time and cost savings. Furthermore, enhanced speed would translate to a competitive advantage, empowering Kreditz's sales team.

H. Using deep learning for streamlined new market entries

Expanding into new markets holds the greatest potential for adding business value to Kreditz, given the demand for bank transaction categorization services and its wide-ranging applications. While Sweden represents a profitable market, its size is relatively small. Leveraging models like MLP and CNN, we can circumvent the need for engineering hard-coded rules and adopt more sophisticated, self-learning approaches. This approach is highly scalable, albeit with a significant caveat: access to labeled training data is essential.

Fortunately, Kreditz already possesses an established pipeline for labeling raw data using consultants in new markets. Consequently, it is conceivable that leveraging consultants for data labeling and subsequently training a DL model would be more cost-effective than engaging them to both label transactions and develop a market-specific engine. This efficiency arises from the inherent scalability of deep learning models, which can autonomously define their own rules without the need for market-specific engineering efforts.

Moreover, compared to engines reliant on a multitude of specific and non-transferable rules, our models offer a more streamlined and adaptable solution. By reducing costs and enhancing profitability for Kreditz, our models create substantial value. They empower the company to swiftly enter new markets, leveraging its existing infrastructure and expertise to capitalize on burgeoning opportunities without the burden of extensive customization and development efforts.

I. Trade-off between precision and recall

For Kreditz, a pivotal decision lies in determining whether to prioritize precision or recall in their predictive engine. Given the specific applications such as income verification, online gambling monitoring, and anti-money laundering efforts, emphasizing higher precision assumes greater importance. A focus on high precision ensures fewer false positives, a critical requirement for these applications. For instance, while it might be tolerable for the model to mistakenly categorize a salary as "others," misclassifying a gift as a salary could have significant repercussions. This distinction is paramount, especially for creditors aiming to safeguard against extending loans to individuals lacking a stable income. Therefore, precision plays a pivotal role in facilitating effective income verification.

The performance of our models, with the exception of BiLSTM, exhibiting high precision across all classes, underscores the efficacy of the MLP, LSTM, and CNN. This affirmation reinforces our confidence in these models' potential to significantly enhance and add value to the categorization engine.

J. Ethical implications

As with any machine learning application, our models are susceptible to biases arising from potential limitations within the training data. One such example relates to the temporal distribution of transactions. If a salary payment deviates from the usual timeframe (e.g. received on the 25th instead of the month's end in Sweden), the model might misclassify it due to a scarcity of training examples reflecting such anomalies. This misclassification could categorize income as "other" or a similar class, potentially leading to negative downstream consequences.

A particularly concerning implication of such misclassification lies in its potential to influence critical decisions such as credit scoring or loan approvals. Inaccuracies in income classification could lead the model to misjudge an individual's creditworthiness, potentially denying access to loans or credit lines despite a history of consistent, substantial income.

K. Cheating the system

Renowned businessman Charlie Munger once asserted, "Show me the incentive, I'll show you the outcome", highlighting the power of incentives in shaping human behavior. In the context of using deep learning models for tasks like credit scoring and anti-money laundering, the presence of incentives invites attempts to exploit system vulnerabilities.

Given the opaque nature of deep learning models, devising robust preventive measures against manipulation is challenging. Transactions are classified based on amount, description, and date, providing opportunities for exploitation. For instance, individuals might manipulate their transactions by altering descriptors to mimic legitimate income sources, such as labeling funds as "lön" (salary) to inflate their creditworthiness. Additionally, there is a risk of disguising illicit money laundering activities as legitimate income by timing transactions to coincide with known payday cycles, such as Sweden's typical salary disbursement date on the 25th of each month.

L. Cost considerations

1) Hardware requirements: Training deep learning models often requires powerful hardware like Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs). Renting cloud computing resources with high-end GPUs/TPUs from providers like Google Cloud, AWS, or Microsoft Azure can be expensive, especially for models with numerous parameters. Training a model like GPT-3, with its 175 billion parameters, is estimated to have cost millions of dollars.

In our case, the proof-of-concept models were relatively small, with only 800-1400 parameters (as seen in Table IV). However, using more extensive data with longer sequences of characters and potentially more complex relationships might require models with parameters in the hundreds of millions, leading to significantly higher hardware costs. Therefore, a cost-benefit analysis considering net present value is crucial when deciding on model complexity.

2) Less is more: Interestingly, our best performing models (CNN and LSTM) were not the ones with the most parameters. Minimizing the number of parameters was a priority across all models to avoid overfitting.

These findings align with existing research. A study by Nikolados et al. (2022) [21] found that while CNNs achieved higher accuracy on protein expression prediction tasks compared to MLPs despite having more parameters, increasing the depth of MLPs only marginally improved their accuracy even with a similar number of parameters. This suggests that the inherent inductive biases of CNNs, which are assumptions the model makes about the data structure, contribute more to their performance than simply the number of parameters.

LSTMs also have inductive biases suited for sequential data, explaining their good performance with a relatively low parameter count. Our results support this, with the CNN and LSTM models performing better than the MLP despite having fewer parameters. The BiLSTM, with the most parameters, had the worst accuracy, further suggesting that a balance between model complexity and data size is essential.

M. Benefits for society

As discussed in our introduction, the research aimed to support several of the Sustainable Development Goals. Upon reviewing our results, we find that our models successfully can contribute to these goals.

Our deep learning models provide an easy way to facilitate access to financial services and markets, supporting goal 9, by more or less automating the credit scoring system. This is especially useful in countries and societies where financial institutions lack sufficient data over their clients.

The models also contribute to reduced inequalities, goal 10, by promoting economic inclusion and that everyone with sufficient credit score should be able to be approved for money lending independent of their age, gender, disability, race, ethnicity, origin or religion.

Our models can also help with responsible consumption and production, which is goal 12. The models can easily be used for tracking various types of transactions, and can help with tracking the overall sustainability of a company. For example, how much a company spends on renewal energy or non-renewable energy can be detected by incorporating these classes into the data.

Last but not least, our models also help with peace, justice and strong institutions around the world. A major use of the models is to detect money laundering, which combats organized crime and illegal financial flows. Abnormalities in the transactions can more easily be detected, and there is a possibility of extending the classes to include a ”money laundering” class for additional security.

VIII. Conclusion

Our research endeavors to harness the power of deep learning technologies to refine transaction categorization accuracy at Kreditz, with a focus on seven strategically significant transaction classes while consolidating less crucial types into an ”others” category. Through rigorous evaluation of various deep learning models, including MLPs, CNNs, LSTMs, and BiLSTMs, we aimed to ascertain their effectiveness in enhancing classification performance.

Despite grappling with the challenges posed by an imbalanced dataset, our models consistently maintained high F1-score for critical classes. Leveraging character-level embeddings proved pivotal in addressing a dataset weakness – succinct and incomplete descriptions. Additionally, the utilization of class weights during training bolstered the models' proficiency in prioritizing minority classes, ensuring a more balanced performance across all categories.

Notably, the LSTM and CNN models attained the highest macro-averaged F1-score, achieving an impressive 97%. Furthermore, considering their efficiency, the MLP and CNN models emerged as faster and less computationally intensive alternatives compared to the LSTM and BiLSTM models, rendering them ideal for real-time or near real-time transaction categorization.

We firmly assert that our models represent the cutting-edge in this domain and possess the capability to categorize transactions with sophistication, thereby imparting substantial value in market expansions.

In summation, our findings underscore the immense potential of deep learning models in augmenting transaction categorization at Kreditz. Beyond enhancing accuracy and speed, they serve as a springboard for new market entries. This research serves as a compelling proof of concept, showcasing how the integration of advanced deep learning techniques can refine and optimize Kreditz's categorization engine, ultimately fostering superior financial insights and decision-making capabilities, ultimately leading to better financial insights and decision-making.

IX. Future work

This research underscores the limitations inherent in addressing the nuanced complexities of tasks like bank transaction categorization using predefined classes. An intriguing avenue for future exploration involves leveraging unsupervised deep clustering techniques, offering the potential to enable new market entries without reliance on labeled data.

Moreover, the integration of Large Language Models (LLMs) holds substantial promise for Kreditz to expand into new geographic markets while minimizing resource allocation. By harnessing its existing engine and focusing primarily on training transaction amounts and dates, Kreditz could streamline its expansion efforts. Additionally, the adoption of multilingual BERT, an encoder-only model utilizing the transformer architecture with proficiency in 104 languages, stands to significantly expedite market entry processes, facilitating global expansion initiatives.

Acknowledgments

The authors would like to thank Kreditz AB for providing access to proprietary transaction data and valuable insights into the financial industry. Their collaboration and support have been instrumental in the success of this research project.

Author Contributions

Filip Orestav is a B.Sc. student in Industrial Engineering and Management at KTH, with a specialization in Computer Science. He conducted a comprehensive literature review, performed rigorous data analysis, was responsible for the vector embeddings and focused on the development of the LSTM and BiLSTM models.

Kolumbus Lindh is a M.Sc. student in Industrial Engineering and Management at KTH, specializing in Machine Learning. His contributions were multifaceted and impactful, encompassing an extensive review of pertinent literature, active involvement in writing, meticulous data analysis, and exploration, with a particular focus on CNNs and MLPs.

REFERENCES

[1] Kreditz, About Kreditz. Retrieved 2024-05-10 from: https://kreditz.com/about

[2] United Nations, Department of Economic and Social Affairs Sustainable Development, THE 17 GOALS. Retrieved 2024-05-10 from: https://sdgs.un.org/goals

[3] United Nations, Department of Economic and Social Affairs Sustainable Development, Goal 12: Ensure sustainable consumption and production patterns. Retrieved 2024-05-10 from: https://sdgs.un.org/goals/goal12s

[4] Kotios, D., Makridis, G., Fatouros, G. et al. Deep learning enhancing banking services: a hybrid transaction classification and cash flow prediction approach. J Big Data 9, 100 (2022). https://doi.org/10.1186/s40537-022-00651-x

[5] Cheikh, F. A., Tarekgn, A. N. & Ullah, M. Deep Learning for Multi-Label Learning: A Comprehensive Survey. arXiv preprint arXiv:2401.16549 (2024). https://arxiv.org/ftp/arxiv/papers/2401/2401.16549.pdf

[6] Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN COMPUT. SCI. 2, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1

[7] Mohammed, A. & Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges, Journal of King Saud University - Computer and Information Sciences, Volume 35, Issue 2, 2023, Pages 757-774, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2023.01.014

[8] Busson, A., Rocha, R., Gaio, R., Miceli, R., Pereira P. I., Moraes, D., Colcher, S., Veiga, A., Rizzi, B., Evangelista, F., Santos, L., Marques, F., Rabaioli, M., Feldberg, D., Mattos, D., Pasqua, J. & Dias, D. (2023). Hierarchical Classification of Financial Transactions Through Context-Fusion of Transformer-based Embeddings and Taxonomy-aware Attention Layer. https://arxiv.org/pdf/2312.07730.pdf

[9] A. Bharambe, A. A. Chandorkar and D. Kalbande. A Deep Learning Approach for Dengue Tweet Classification. Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 1043-1047, doi: 10.1109/ICIRCA51532.2021.9544862. https://ieeexplore.ieee.org/document/9544862

[10] Xia, F. Label Oriented Hierarchical Attention Neural Network for Short Text Classification. Academic Journal of Engineering and Technology Science (2022) Vol. 5, Issue 1: 53-62. https://doi.org/10.25236/AJETS.2022.050111. https://doi.org/10.25236/AJETS.2022.050111

[11] Lipton, Z.C., Berkowitz, J. and Elkan, C. A critical review of recurrent neural networks for sequence learning, arXiv preprint arXiv:1506.00019, 2015. https://arxiv.org/abs/1506.00019

[12] Hochreiter, S. and Schmidhuber, J. Long short-term memory, *Neural Computation*, 9(8), 1997, pp. 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735

[13] Huang, J., Chai, J. & Cho, S. Deep learning in finance and banking: A literature review and classification. Front. Bus. Res. China 14, 13 (2020). https://doi.org/10.1186/s11782-020-00082-6

[14] Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen & Y, Sun, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support Systems, Volume 50, Issue 3, 2011, Pages 559-569, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2010.08.006.

[15] The Nilson Report, *Card Fraud Worldwide*, 2023.

[16] Agarwal, S., Alok, S., Ghosh, P., & Gupta, S. Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech (December 21, 2019). Indian School of Business. http://dx.doi.org/10.2139/ssrn.3507827

[17] A. Demirgüç-Kunt, L. Klapper, D. Singer, S. Ansar, and J. Hess, The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution, The World Bank, 2018.

[18] Chen, Z., Van Khoa, L.D., Teoh, E.N. et al. Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. Knowl Inf Syst 57, 245–285 (2018). https://doi.org/10.1007/s10115-017-1144-z

[19] Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X. & Ward, R. Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. 2015. https://arxiv.org/abs/1502.06922

[20] Verified Market Research, Financial Reporting Software Market. Retrieved 2024-05-10 from: https://www.verifiedmarketresearch.com/product/financial-reporting-software-market/

[21] Nikolados, EM., Wongprommoon, A., Aodha, O.M. et al. Accuracy and data efficiency in deep learning models of protein expression. Nat Commun 13, 7755 (2022). https://doi.org/10.1038/s41467-022-34902-5

TRITA-EECS-EX-2024:274
Stockholm, Sverige 2024

www.kth.se