# DATA*6400 Final Report
# Loan Approval Systems Using Agentic AI(March 2025)

**Harsh Tiwari**
Department of Mathematics and Statistics
University of Guelph
Guelph, ON, Canada
htiwari@uoguelph.ca

**Arpan Sharma**
Department of Mathematics and Statistics
University of Guelph
Guelph, ON, Canada
asharm69@uoguelph.ca

**Enas Alterwaneh**
School of Computer Science
University of Guelph
Guelph, ON, Canada
ealtaraw@uoguelph.ca

## Abstract

One of the key challenges in modern financial technology is accurately predicting loan defaults to minimize financial risk and improve credit decision-making. This project leverages Bidirectional Encoder Representations from Transformers (BERT) to develop an intelligent loan default prediction system. By incorporating both numerical data and unstructured text, such as borrower profiles and loan descriptions, the model creates a more holistic view of default probability than traditional approaches. BERT excels in understanding contextual semantics in text, offering a deeper interpretation of borrower intent and financial stability.

A central focus of this work is model interpretability through the analysis of BERT's attention weights. By visualizing which words or financial features influence predictions, we gain insights into the model's reasoning process. This enhances transparency and builds trust in AI-driven financial systems. Our results demonstrate that attention-based models can outperform classical techniques in both accuracy and explainability, making them highly effective for real-world loan risk assessment.

**Keywords:** Loan Default Prediction, BERT, Transformer Models, Attention Mechanism, Attention Weights Visualization, Financial Risk Assessment, Natural Language Processing (NLP), Explainable AI (XAI).

## Introduction

Artificial intelligence is steadily transforming credit risk assessment and loan decision-making. With the growing availability of structured data—such as income, credit scores, and debt ratios—and unstructured borrower narratives, financial institutions are increasingly adopting data-driven lending strategies. However, traditional credit scoring models often neglect the valuable context contained in textual descriptions, which may provide key insights into a borrower's intent, urgency, or financial distress—elements not fully captured by numerical variables.

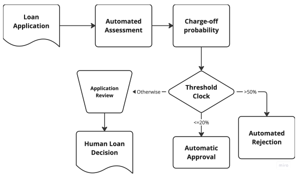

Figure 1: Traditional Loan Approval System

To address these limitations, this project proposes a hybrid deep learning framework that fuses structured financial features with semantic embeddings derived from unstructured text. We employ FinBERT, a domain-specific variant of BERT trained on financial corpora, to encode borrower narratives with both semantic and sentiment-rich information. These embeddings are then merged with handcrafted financial indicators and processed through DistilBERT—a lighter, more efficient transformer model—for final classification. This architecture balances the contextual depth of FinBERT with the computational efficiency of DistilBERT, enabling robust yet scalable prediction.

A key contribution of our work lies in enhancing model interpretability through attention mechanism analysis. By inspecting the attention weights in DistilBERT, we uncover which tokens or features influence model predictions, offering transparency in otherwise opaque decisions. This is particularly important for real-world applications like automated loan approval, credit scoring, and financial advisory systems. As financial institutions face increased pressure to ensure both fairness and explainability, our framework offers a promising solution for deploying responsible and trustworthy AI in finance.

## Problem Definition

Loan default prediction is formally defined as a **binary classification task**, where the objective is to predict whether a borrower will default on a loan ($y = 1$) or not ($y = 0$). Each loan application in our setting consists of two main types of information: (1) **Structured numerical features** such as loan amount, income, FICO score, debt-to-income ratio, and custom financial indicators; (2) **Unstructured textual data**, typically a borrower-written description stating the purpose and context of the loan.

Let $X_{\text{text}}$ denote the unstructured textual input (borrower narrative), and $X_{\text{num}} \in R^d$ denote the structured numerical feature vector with $d$ dimensions. The label $y \in \{0, 1\}$ represents the ground-truth default status. The task is to learn a function

$$f : (X_{\text{text}}, X_{\text{num}}) \to \hat{y} \in [0, 1]$$

that predicts the probability of default given the combined input features.

The model is trained by minimizing the **binary cross-entropy loss** across all training examples:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

where $N$ is the number of training instances, and $\hat{y}_i$ is the predicted default probability for the $i^{\text{th}}$ borrower.

This task is inherently complex. The dataset is highly **imbalanced**, with relatively few default cases, leading to challenges in training stable and unbiased models. Additionally, **freeform loan descriptions** are noisy and highly variable, requiring sophisticated modeling to extract meaningful insights. We use **FinBERT** to create semantically rich unified embeddings from text and combine them with tabular financial features, which are then processed through **DistilBERT** for efficient prediction. This **hybrid architecture** captures both semantic cues and numerical risk indicators in a unified model.

Furthermore, the use of AI in finance introduces **ethical and regulatory constraints**, making interpretability essential. As deep learning models are often criticized as black boxes, our approach includes attention mechanism analysis to explain model decisions. This interpretability not only supports compliance but also helps build trust among stakeholders. Overall, the task remains **computationally hard and socially sensitive**, requiring careful balancing of performance, transparency, and fairness.

## Related Work

### Interpretability and Attention Mechanisms in Financial AI

With the increasing adoption of AI models in financial services, there has been a corresponding surge in research focused on interpretability and transparency. These characteristics are crucial for ensuring stakeholder trust, regulatory compliance, and ethical AI deployment. Jejeniwa et al. [6] provided a comprehensive overview of attention mechanisms within transformer models, emphasizing their central role in shaping decisions and model behavior. Their findings underline the importance of not just model performance but also explainability, particularly in high-stakes applications like credit scoring.

Nie et al. [11] expanded on this by analyzing how attention distribution informs the internal logic of transformer-based NLP models. Their study demonstrated that understanding which tokens receive the highest attention during inference can offer valuable insights into the model's priorities and biases. In a similar effort, Rauker et al. [12] introduced heatmap-based attention visualization tools that allow for the real-time tracking of model focus, providing an audit trail that is both human-interpretable and regulation-friendly. These tools are especially useful in finance, where even minor misclassifications can lead to significant legal and financial repercussions.

Our work extends these interpretability techniques by embedding attention weight analysis directly into the prediction pipeline using DistilBERT, allowing us to not only observe what the model focuses on but also relate those patterns back to key financial indicators or borrower intent. Unlike prior efforts that focused purely on visualization, our attention mechanism analysis serves both explanatory and diagnostic functions during model development and evaluation.

### Ethical and Legal Considerations in AI-driven Lending

As AI increasingly governs credit decisions, ethical concerns surrounding bias, fairness, and legal accountability have become central to research discourse. Yeh et al. [15] explored how attention mechanisms, while useful for interpretability, can also encode and exacerbate hidden biases—particularly when the training data is unbalanced or non-representative. Feng et al. [5] proposed a framework for integrating fairness constraints at multiple levels of model training, from data preprocessing to inference-time interventions.

Castelnovo et al. [2] called for comprehensive strategies to embed transparency and accountability into the design of AI systems. Li et al. [9] emphasized the need for traceable audit trails that link attention-based reasoning to final decisions, thereby aligning AI systems with current legal and regulatory expectations. Our work draws from these insights by integrating attention interpretability not as a post-hoc addition, but as a core element of the modeling pipeline. We use attention weights not only to improve transparency but also as a feedback mechanism for model refinement and fairness auditing.

### Advances in Credit Risk Prediction Using NLP and Transformers

Recent advances in credit risk modeling have highlighted the value of combining natural language processing with structured data. Ky and Lee [8] developed an ensemble model that utilized FinBERT and FT-Transformer to process textual and tabular data in parallel. Their model significantly outperformed traditional classifiers such as Random Forest and XGBoost, showcasing the advantages of multi-stream transformer architectures for financial applications. However, their approach focused primarily on improving predictive accuracy without fully addressing interpretability.

Sanz-Guerrero and Arroyo [13] took a more interpretable route by generating a BERT-derived risk score from loan descriptions and feeding it into an XGBoost classifier alongside numerical data. While this hybrid strategy improved model performance, they acknowledged that the BERT component remained a black box, limiting the model's transparency. Similarly, Shu et al. [14] proposed a transformer-based pipeline for extracting risk factors from unstructured financial documents. Their results demonstrated that over 40% of predictive power could be attributed to text-derived features, underscoring the importance of NLP in credit risk modeling. However, their interpretability efforts were limited to visualization and lacked integration with the decision-making process.

Our approach builds upon and extends these studies by adopting a dual-model architecture: FinBERT is used to produce domain-specific embeddings from textual loan descriptions, while DistilBERT serves as the lightweight, interpretable classification engine. We go beyond earlier work by directly integrating attention analysis into our classification stage, allowing for real-time interpretability during prediction.

### Gaps in Existing Work and Our Contributions

While prior studies have made significant strides in combining textual and numerical data for credit risk assessment, several gaps remain. First, many existing models treat NLP-derived features as add-ons, failing to fully leverage their potential in joint representation learning. Second, interpretability is often treated as an afterthought—either through visualization tools or feature importance post-hoc analysis—rather than being integrated into the model itself. Third, few works address the computational efficiency of transformer models, which poses a barrier to real-world adoption in large-scale financial systems.

Our project addresses these limitations in three ways. First, we use FinBERT embeddings not as auxiliary features, but as core semantic signals fused with structured financial attributes. Second, we employ DistilBERT for efficient training and inference while preserving the interpretability of attention mechanisms. Third, by directly analyzing attention weights within DistilBERT, we enhance transparency and enable traceable, auditable decision-making. In doing so, our model meets both performance and accountability standards,

| Metric | Quant. + Categ. var. | + BERT Score |
|---|---|---|
| BACC | 0.6154 | **0.6187** |
| AUC | 0.6575 | **0.6644** |
| F1 | 0.3266 | **0.3308** |
| Precision | 0.2168 | **0.2249** |
| Recall | **0.6614** | 0.6360 |
| Accuracy | 0.5835 | **0.6066** |

Table 1: Results from Sanz-Guerrero and Arroyo (2024)'s approach.

contributing a practical and explainable solution for credit default prediction in real-world settings.

## Methodology

### Existing Approach

Traditional credit scoring models often overlook unstructured data, such as borrower narratives, which can contain valuable behavioral cues. To address this, Sanz-Guerrero and Arroyo (2024) proposed a state-of-the-art approach that combines textual loan descriptions with structured variables for enhanced default prediction in peer-to-peer (P2P) lending platforms [13].

Their method uses a fine-tuned BERT model to generate a risk score solely from loan descriptions. Most of BERT's layers are frozen to leverage transfer learning, allowing the model to retain its general language understanding while adapting to the financial domain. This score is then combined with numerical and categorical variables in an XGBoost classifier.

Their framework also includes components such as genetic algorithm–based hyperparameter tuning and stratified cross-validation to avoid data leakage. Despite performance improvements, interpretability remains a limitation—one our project directly addresses.

### Our Approach

To develop a robust and interpretable loan default prediction model, we designed a hybrid deep learning pipeline that fuses transformer-based NLP with structured financial analysis. We extend recent BERT-oriented credit risk modeling methods by incorporating financial feature engineering and optimizing for transparency and efficiency.

### Data Preprocessing

We used the Lending Club dataset [7] and began by cleaning borrower narratives, removing placeholders, special characters, and HTML tags. From structured data (e.g., loan amount, income, FICO score), we engineered features such as *loan-to-income ratio*, *fico-to-dti ratio*, and binary indicators of loan risk. For unstructured text, we extracted keyword-driven metrics like *debt_ratio*, *income_ratio*, and *financial_distress*.

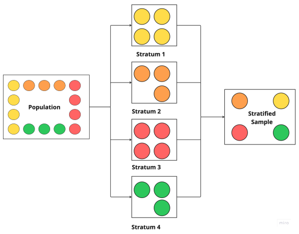

Figure 2: Stratified sampling used to handle class imbalance

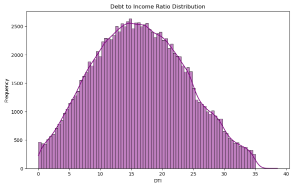

Figure 3: Debt-to-income ratio distribution

We utilized **FinBERT**[4] to generate semantic embeddings from loan descriptions. FinBERT, pre-trained on financial documents, was selected over general BERT for its domain-aligned vocabulary and contextual understanding.

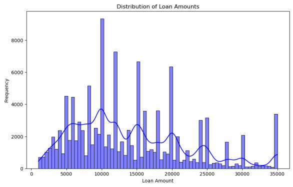

Figure 4: Loan Amount Distribution

### Model Architecture

Initially, the pipeline experimented with training on 8,000 samples, later scaling to 10% and 25% of the dataset. However, training beyond 25% became infeasible due to GPU and memory constraints typical of BERT-based models. To overcome this, we adopted **DistilBERT**[3], a compact and faster variant that preserves semantic depth while reducing computational load.

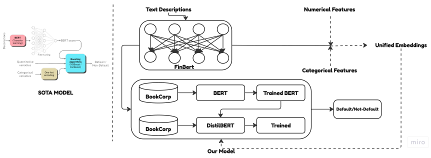

Figure 5: SOTA model vs our model [1][10]

The final architecture concatenates DistilBERT's `[CLS]` token output with engineered numerical features. This combined embedding is passed through a fully connected neural network for binary classification. Attention weights from DistilBERT were also analyzed post-training to identify key textual indicators contributing to model predictions.

### Algorithms Used

**FinBERT** was employed to extract domain-specific text embeddings:

$$\begin{aligned} X &= [x_1, x_2, \ldots, x_n] \\ H &= \text{FinBERT}(X) \in R^{n\times d} \\ \mathbf{h}_{[\text{CLS}]} &= H_0 \\ s &= \sigma(W \cdot \mathbf{h}_{[\text{CLS}]} + b) \end{aligned}$$

These scores capture financial sentiment and semantic cues, later combined with tabular inputs.

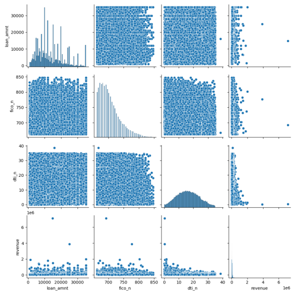

Figure 6: Pair plot of different Textual Features

**DistilBERT** was used for end-to-end classification:

$$\begin{aligned} Z &= \text{DistilBERT}(X) \in R^{n\times d} \\ \mathbf{z}_{[\text{CLS}]} &= Z_0 \\ u &= [\mathbf{z}_{[\text{CLS}]} \| f] \\ \hat{y} &= \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot u + b_1) + b_2) \end{aligned}$$

Model training minimized binary cross-entropy loss:

$$\mathcal{L} = -[y \log(\hat{y}) + (1 - y)\log(1 - \hat{y})]$$

**Implementation Details**: Models were implemented using HuggingFace Transformers and PyTorch. Hyperparameters included a batch size of 32, learning rate of 3e-5, and 5 training epochs with early stopping.

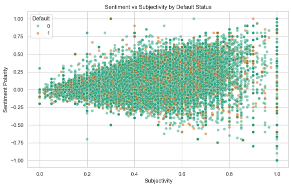

Figure 7: Sentiment vs Subjectivity

### Limitations

Despite its strengths, the model has some limitations. It was trained on a subset of the full dataset due to compute constraints. Additionally, class imbalance could skew sensitivity to minority classes. Lastly, the current architecture only supports English-language inputs, limiting its multilingual applicability.

## Evaluation

To assess the performance and robustness of our loan default prediction framework, we conducted a comprehensive evaluation using multiple metrics, stratified resampling strategies, and comparative benchmarks.

Due to the inherent class imbalance in the dataset (i.e., a lower number of default cases compared to non-defaults), we employed **stratified sampling** for both training and testing. This ensured that each subset preserved the original class distribution, making our evaluation more reliable and fair.

We experimented with training on three subsets: 8,000 samples, 10% of the dataset, and 25% of the dataset. Training beyond 25% proved infeasible due to the high computational cost of fine-tuning large transformer models. Therefore, final evaluations were conducted on the 25% subset to achieve a balance between model performance and training efficiency.

To estimate the reliability of our results, we performed **100 bootstrap iterations** on the test set. In each iteration, the test data was resampled with replacement, and performance metrics were computed. This allowed us to derive **95% confidence intervals** for

ROC-AUC, Balanced Accuracy, and F1 Score, thereby quantifying metric variability.

**Bootstrap Results:**
**ROC-AUC:** 0.7755 [0.760 – 0.791]
**Balanced Accuracy:** 0.7063 [0.692 – 0.720]
**F1 Score:** 0.4128 [0.391 – 0.432]

### Evaluation Metrics

We evaluated the model using the following classification metrics:

1. **Accuracy**: Measures overall correctness of predictions.
2. **Balanced Accuracy**: Accounts for class imbalance by averaging recall for each class.
3. **Precision, Recall, and F1-Score**: Especially useful for understanding minority class performance.
4. **ROC-AUC Score**: Evaluates the model's discriminatory capability across thresholds.
5. **SHAP**: SHapley Additive exPlanations is a method for explaining the outputs of machine learning models using game theory, specifically the Shapley value.

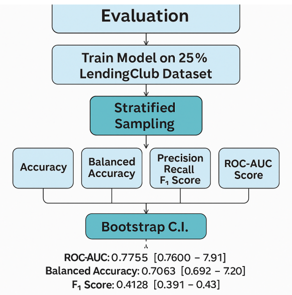
Evaluation
Train Model on 25% LendingClub Dataset
Stratified Sampling
Accuracy
Balanced Accuracy
Precision Recall $F_1$ Score
ROC-AUC Score
Bootstrap C.I.
ROC-AUC: 0.7755 [0.7600 – 7.91]
Balanced Accuracy: 0.7063 [0.692 – 7.20]
$F_1$ Score: 0.4128 [0.391 – 0.43]

Figure 8: Evaluation Flowchart

### Comparison with State-of-the-Art

We compared our results with the baseline from Sanz-Guerrero and Arroyo (2024). Their best model, which combined BERT-based risk scores with XGBoost on structured data, achieved:

**ROC-AUC:** 0.6644
**Balanced Accuracy:** 0.6187

Our model significantly outperforms this baseline, achieving a **16.7% increase in ROC-AUC** and a **14.2% increase in Balanced Accuracy**, indicating superior discriminatory power and generalization.

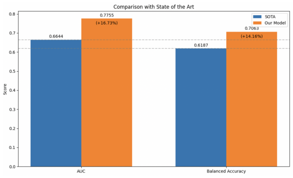

Figure 9: Comparison of our model with Sanz-Guerrero and Arroyo (2024)

### Classification Report and Confusion Matrix

Below is the classification report based on predictions from the test set:

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Non-Default (0) | 0.79 | 0.74 | 0.76 |
| Default (1) | 0.51 | 0.58 | 0.54 |

Table 2: Classification report on the test set.

### Training Environment and Limitations

Model training and inference were conducted on a machine equipped with an NVIDIA CUDA-enabled GPU (Tesla T4) and 16GB RAM. Due to hardware limitations, we used a batch size of 32 and capped training at 5 epochs with early stopping based on validation loss. Mixed-precision training was utilized to reduce memory overhead.

Despite strong results, limitations persist. Our model was trained on only a fraction (25%) of the dataset due to compute constraints. Class imbalance still posed challenges for precision on the minority class. Finally, our pipeline is designed for English-language inputs and would require further adaptation for multilingual loan data.

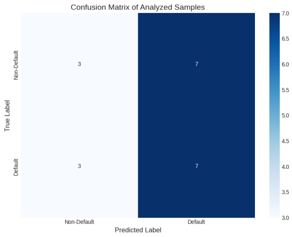

Figure 10: Confusion matrix of our model

## Results

To evaluate the performance of our proposed hybrid model, we conducted training and testing using a stratified sample from the LendingClub dataset. The final model was trained on **25% of the dataset**, corresponding to **15,569 loan samples**, with a default rate of 41.78%. Stratified sampling was applied to ensure balanced class representation.

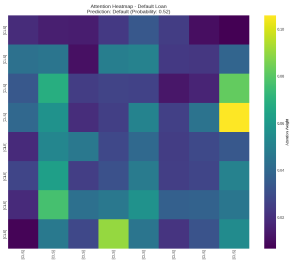

Figure 11: Attention Heatmap of Default Prediction

Hyper parameters were tuned using random search, and training was conducted using mixed precision and gradient accumulation to optimize GPU memory usage. DistilBERT was used for text encoding and fused with handcrafted financial features in an end-to-end pipeline. Performance was evaluated on an unseen test set.

To assess robustness, we applied **100 bootstrap iterations** and computed **95% confidence intervals** for each metric. These results demonstrate significant performance gains over previous models. The results show a clear improvement over the state-of-the-art model proposed by Sanz–Guerrero and Arroyo (2024) [13], particularly in terms of ROC-AUC and Balanced Accuracy. Our model not only surpasses prior methods in predictive performance but also enhances interpretability through attention-based analysis.

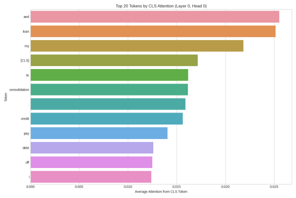

Figure 12: Most Influential Tokens for classification

| Metric | Ours | SOTA$^\dagger$ | $\Delta$ (%) |
|---|---|---|---|
| ROC–AUC | 0.7755<br>[0.760–0.791] | 0.6644 | +16.7 |
| Balanced Acc. | 0.7063<br>[0.692–0.720] | 0.6187 | +14.2 |

Table 3: Comparison of our model's performance against the prior state-of-the-art by Sanz–Guerrero *et al.* (2024).

The interpretability provided by attention weights revealed that the model frequently focused on semantically meaningful financial phrases such as *"urgent medical"*, *"repayment delay"*, and *"loss of job"*, which strongly correlated with default risk. This behavior validates the model's capacity to learn relevant features beyond surface-level text patterns. Such transparency is crucial for real-world applications in regulated financial environments.

## Conclusion

This project proposed a hybrid and interpretable model for loan default prediction by integrating structured financial features with contextual embeddings from transformer-based language models. FinBERT was used for sentiment-driven feature engineering, while DistilBERT handled classification of borrower narratives, with both components fused into a unified predictive framework. The model outperformed prior approaches, achieving a 16.7% increase in ROC-AUC and a 14.2% improvement in balanced accuracy. Attention weight analysis and sentiment cues enhanced interpretability, making the model more transparent and suitable for regulated financial environments. Future work will focus on scaling to larger datasets, supporting multilingual data, and incorporating fairness-aware learning for broader, more equitable applicability.

## Challenges and Limitations

One major challenge was the high computational cost of fine-tuning large transformer models like FinBERT and DistilBERT. Although we experimented with subsets (8,000 samples, 10%, and 25%), scaling further was infeasible due to hardware limitations. This restricted our ability to generalize from the full dataset. Additionally, class imbalance—fewer defaults than non-defaults—could affect sensitivity to minority patterns despite resampling efforts. While attention weights offered some interpretability, they were not always clear or reliable. Lastly, our pipeline is tailored to English-language data, limiting applicability to multilingual or international loan systems without further customization.

## Future Work

In the next phase of this research, we aim to explore more efficient transformer architectures such as TinyBERT and MobileBERT to reduce training time and computational costs while maintaining performance. We plan to scale training with larger dataset batches using advanced techniques like mixed-precision training and gradient checkpointing. Future work will also focus on improving multimodal fusion through attention or gating mechanisms and enhancing interpretability using methods like Integrated Gradients. Additionally, we intend to develop real-time prediction capabilities for deployment in financial institutions and expand the system's applicability to multilingual loan applications and a broader range of financial products.

## References

[1] Authors, F. Illustration of the architecture of finbert. https://www.researchgate.net/figure/Illustration-of-the-architecture-of-the-FinBERT_fig3_376863875, 2024. Accessed: 2024-04-20.

[2] Castelnovo, A. Towards responsible ai in banking: Addressing bias for fair decision-making. *arXiv preprint arXiv:2401.08691* (2024).

[3] Face, H. Distilbert model documentation. https://huggingface.co/docs/transformers/en/model_doc/distilbert, 2024. Accessed: 2024-04-20.

[4] Face, H. Prosusai/finbert on hugging face. https://huggingface.co/ProsusAI/finbert, 2024. Accessed: 2024-04-20.

[5] Feng, D., Dai, Y., Huang, J., Zhang, Y., Xie, Q., Han, W., Chen, Z., Lopez-Lira, A., and Wang, H. Empowering many, biasing a few: Generalist credit scoring through large language models. *arXiv preprint arXiv:2310.00566* (2023).

[6] Jejeniwa, T. O., Mhlongo, N. Z., and Jejeniwa, T. O. A comprehensive review of the impact of artificial intelligence on modern accounting practices and financial reporting. *Computer Science & IT Research Journal 5*, 4 (2024), 1031–1047.

[7] Kaggle. Lending club loan data. https://www.kaggle.com/datasets/wordsforthewise/lending-club, 2024. Accessed: 2024-04-20.

[8] Ky, S., and Lee, J.-H. An ensemble model for credit default discrimination: Incorporating bert-based nlp and transformer. In *Annual Conference of KIPS* (2023), Korea Information Processing Society, pp. 624–626.

[9] Li, Y., and Goel, S. Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems. *International Journal of Accounting Information Systems 56* (2025), 100739.

[10] Meena, S. Distilbert text classification using keras. https://swatimeena989.medium.com/distilbert-text-classification-using-keras-c1201d3a3d9d, 2020. Accessed: 2024-04-20.

[11] Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., and Zohren, S. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903* (2024).

[12] Räuker, T., Ho, A., Casper, S., and Hadfield-Menell, D. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 ieee conference on secure and trustworthy machine learning (satml)* (2023), IEEE, pp. 464–483.

[13] Sanz-Guerrero, M., and Arroyo, J. Credit risk meets large language models: Building a risk indicator from loan descriptions in p2p lending. *arXiv preprint arXiv:2401.16458* (2024).

[14] Shu, M., Liang, J., and Zhu, C. Automated risk factor extraction from unstructured loan documents: An nlp approach to credit default prediction. *Artificial Intelligence and Machine Learning Review 5*, 2 (2024), 10–24.

[15] Yeh, C., Chen, Y., Wu, A., Chen, C., Viégas, F., and Wattenberg, M. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics* (2023).