Mitigating Adversarial Attacks on Transformer Models in Credit Scoring

Brandon Schwab[1, *] and Johannes Kriebel[2, †]

[1]Institute for Risk and Insurance, Leibniz University Hannover, Hannover, Germany
[2]Faculty of Business, Economics and Social Sciences, University of Hamburg, Hamburg, Germany
[*]Corresponding author. Address: Königsworther Platz 1, 30167 Hannover, Germany. E-mail: brandon.schwab@insurance.uni-hannover.de
[†]Sadly, Johannes Kriebel passed away in February 2025. As this work was originally co-authored with him, it is submitted in recognition of his contribution and in his memory.

This version: March 30, 2025

Abstract

The integration of unstructured data, such as text created by borrowers, offers new opportunities for improving credit default prediction but also introduces new risks. This study examines the robustness of transformer-based credit scoring models that utilize textual data and assesses their vulnerability to adversarial attacks. Using peer-to-peer lending data, we show that small, semantically neutral changes in loan descriptions can substantially alter model outputs. These vulnerabilities expose lenders and borrowers to economic risks through distorted risk assessments and mispriced loans. We evaluate two mitigation strategies: adversarial training and topic modeling. Adversarial training improves robustness without compromising predictive performance. Topic modeling provides a more interpretable and stable representation of borrower narratives. An economic analysis confirms that robust models reduce mispricing and improve outcomes for all parties. The findings underscore the importance of robustness as the use of unstructured data in credit scoring becomes more accessible.

Keywords: OR in banking, Peer-to-peer lending, Deep learning, Textual data, Credit risk

[0]We thank Judith Schneider for her valuable comments and suggestions.

1 Introduction

Credit default prediction is a critical field within financial risk management with significant implications for lenders, borrowers, and the broader financial system. Traditionally, credit scoring models have relied primarily on structured data, such as financial ratios, borrower demographics, and credit history (Baesens et al., 2003; Crook et al., 2007). The advent of machine learning models, including decision trees, random forests, gradient boosting, and neural networks, has further improved predictive accuracy by capturing complex, nonlinear relationships within structured data (Crook et al., 2007; Lessmann et al., 2015). Ensemble methods, in particular, have consistently outperformed simpler approaches in this domain (Dumitrescu et al., 2022; Fitzpatrick & Mues, 2016; Gunnarsson et al., 2021; Lessmann et al., 2015; Xia et al., 2017).

With the rise of digital technologies, the availability of unstructured data—such as text from loan applications, digital footprints, and social media activity—has expanded, offering new information for credit risk assessment. Recent studies demonstrate that such unstructured data can contain important credit-relevant signals, thereby enhancing the accuracy of credit default predictions substantially (Berg et al., 2020; Iyer et al., 2016; Lin et al., 2013; Tsai & Wang, 2017). Prior literature has first addressed the challenge of utilizing this unstructured data through various information extraction methods, including the identification of specific textual features such as the frequency of identity claims, readability, and sentiment, which are linked to borrower quality (Chen et al., 2018; Dorfleitner et al., 2016; Herzenstein et al., 2011). More recently, deep learning techniques, including modern transformer models, have shown promising results in automatically processing and analyzing unstructured data from text for credit scoring (Ahmadi et al., 2018; Borchert et al., 2023; Fitzpatrick & Mues, 2021; Kriebel & Stitz, 2022; Mai et al., 2019; Matin et al., 2019; Stevenson et al., 2021; Wu et al., 2023; Yu et al., 2024).

Despite these advances, growing attention in artificial intelligence research has focused on the robustness of deep learning models. In particular, such models have been shown to be vulnerable to adversarial attacks—small, often imperceptible changes to input data that can lead to significant shifts in model predictions (Barreno et al., 2010; Goodfellow et al., 2014; Szegedy et al., 2014). While much of this work has focused on image data, recent studies have also revealed similar vulnerabilities in natural language processing tasks, including sentiment analysis, text classification, and named entity recognition (Alzantot et al., 2018; Wang et al., 2019; Zang et al., 2019). These findings raise serious concerns for high-stakes applications like credit scoring, where even subtle textual variations in loan applications could lead to significant changes in predicted risk. Although adversarial robustness has received increasing attention in natural language processing, its implications for credit scoring models that integrate unstructured borrower narratives remain largely unexplored.

This paper addresses that gap. We provide a systematic assessment of the robustness of credit scoring models that incorporate unstructured text data. We demonstrate that even subtle changes in borrower-provided narratives—such as the replacement of a single synonym—can significantly alter model predictions. These findings underscore a critical risk: while textual data can improve predictive performance, it also opens the door to strategic manipulation and unintended disparities, where borrowers may be unfairly penalized due to innocuous linguistic variation.

To address this challenge, we make five main contributions. First, we identify and analyze the vulnerabilities of credit default prediction models that incorporate unstructured data to adversarial attacks. Using peer-to-peer lending data, including textual descriptions provided by borrowers alongside structured financial information, we evaluate manipulated text inputs and assess the extent to which these models can be misled into making inaccurate predictions.

Second, we use explainable artificial intelligence conducting a word attribution analysis at the sentence level to assess the vulnerabilities in more detail. By analyzing how even subtle changes in word choice alter the model's output substantially, we show how a specific choice of words can disproportionately influence risk predictions. This reveals model sensitivities and biases that may affect fairness in credit scoring.

Third, we examine adversarial training as a potential defense strategy. By retraining models on manipulated examples, we find that this approach substantially improves robustness, mitigating the effects of adversarial inputs and preserving performance under attack.

Fourth, we explore topic modeling as an alternative textual representation that may offer greater inherent stability. Instead of relying on fine-grained word embeddings, this approach condenses borrower narratives into high-level thematic structures. We find that topic-based models are significantly less sensitive to adversarial perturbations, suggesting that abstraction away from specific word choices can enhance robustness without sacrificing interpretability.

Fifth, we assess the economic implications of adversarial vulnerabilities by linking shifts in predicted default probabilities to changes in loan pricing. Our results demonstrate that adversarial attacks can lead to substantial misjudgments in loan conditions offered to customers. Interestingly, we show that this could not only disadvantage lenders but often also borrowers.[1] Through this analysis, we highlight the real-world consequences of adversarial vulnerabilities and the importance of adopting robust models to minimize economic risks for lenders and borrowers alike.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature on credit default prediction and the integration of unstructured data. Section 3 outlines our data sources and preprocessing steps. Section 4 describes our methodology for fine-tuning BERT models, generating adversarial samples, assessing model robustness, applying adversarial training, and implementing topic modeling. Section 5 presents the empirical results, including the impact of adversarial attacks on model performance. Section 6 discusses the economic implications of adversarial attacks, and Section 7 concludes the study.

2 Related Literature

The task of predicting credit default has long relied on structured data, such as financial ratios, credit scores, and borrower demographics. Foundational studies, including those by Baesens et al. (2003) and Kumar and Ravi (2007), systematically compared classification algorithms and demonstrated that more complex models, such as neural networks, often outperform traditional regression-based approaches. Subsequent research has reinforced these findings. Studies by Lessmann et al. (2015), Fitzpatrick and Mues (2016), Xia et al. (2017), Gunnarsson et al. (2021), and Dumitrescu et al. (2022) emphasize the strong predictive performance of ensemble methods, particularly random forests and gradient boosting. More recently, methodological advances have included graph representation learning (Shi et al., 2024) and the application of transformer-based models to multi-horizon default prediction (Korangi et al., 2023).

Building on these foundations, more recent research has expanded to include unstructured data, recognizing its potential to further improve credit default prediction. Unstructured data sources, such as text, images, and social media activity, contain credit-relevant information that is often not captured by structured data. Lin et al. (2013) analyze online friendships in peer-to-peer lending networks and reveal that these social connections serve as strong indicators of creditworthiness, effectively reducing information asymmetry and improving default prediction. Iyer et al. (2016) investigate peer assessments on peer-to-peer lending platforms, finding them to be more predictive of default than

[1] The results are, therefore, also of high importance for the debate on fairness in credit scoring (Kozodoi et al., 2022).

conventional credit scores. Similarly, Óskarsdóttir et al. (2019) integrate mobile phone data and social network analytics with structured data, yielding significant improvements in predictive accuracy. Berg et al. (2020) demonstrate that digital footprints, such as device type and online behavior, can not only match but also complement traditional credit bureau scores in predicting defaults. Zandi et al. (2025) use the geographic proximity and information on mortgage providers to include borrower connections and use this to enhance probability of default predictions. With respect to text data in credit scoring, several studies have shown that borrower-provided narratives can contain predictive signals related to default risk. Herzenstein et al. (2011) find that identity claims in loan application in peer-to-peer lending increase the likelihood of funding but are associated with poorer loan performance, including a higher likelihood of default or late payment. Dorfleitner et al. (2016) link textual characteristics such as spelling errors, text length, and positive keywords to successful funding, though their influence on default risk is less clear. Chen et al. (2018) show that excessive punctuation reduces readability, negatively affecting both funding and default rates. Gao et al. (2018) find that clearer, more positively worded loan descriptions are correlated with lower default rates. Tsai and Wang (2017) demonstrate a strong correlation between financial sentiment words in texts and financial risk, while Agarwal et al. (2016) emphasize the predictive value of linguistic tone in credit rating reports, revealing the role of sentiment in assessing credit risk.

Beyond identifying predictive patterns in text, a growing number of studies have focused on systematically extracting and operationalizing such information through natural language processing techniques. Netzer et al. (2019) and Xia et al. (2020) employ methods such as term frequency-inverse document frequency (TF-IDF) and topic modeling to convert textual data into structured features for use in predictive models. Jiang et al. (2018) similarly apply latent Dirichlet allocation to loan descriptions, showing that topic-based features improve default prediction when combined with structured data. Fitzpatrick and Mues (2021) utilize biterm topic modeling to extract features from short loan texts, effectively addressing challenges posed by brevity in peer-to-peer lending narratives. In a related approach, Wang et al. (2020) propose a method for identifying semantic soft factors by mapping loan texts into an embedding space and clustering terms into semantic cliques.

Building on these approaches, more recent studies apply deep learning techniques to further enhance the processing of textual data in credit risk assessment. Ahmadi et al. (2018) use dependency sensitive convolutional neural networks combined with sentiment analysis to detect signs of financial distress in business reports, effectively identifying early indicators of bankruptcy. Matin et al. (2019)

utilize a convolutional recurrent neural network to analyze management statements and auditor reports, achieving notable improvements in predictive accuracy for large firms. Similarly, Mai et al. (2019) evaluate both average embedding models and convolutional neural networks for bankruptcy prediction based on textual disclosures, finding that these methods, particularly average embedding models, enhance predictive accuracy.

A further advancement in modeling textual data for credit risk assessment involves the use of transformer-based architectures. Stevenson et al. (2021) apply BERT to the task of predicting small business loan defaults. Their study shows that while BERT significantly improves predictive performance when used alone, combining it with structured data does not yield additional improvements in their data. Kriebel and Stitz (2022) conduct a benchmark comparison of various deep learning models, including transformers, in the context of peer-to-peer lending. They find that integrating text-based predictions with structured features leads to a significant improvement in overall model performance. Interestingly, they also report that simpler architectures, such as average embedding neural networks, can perform comparably to more complex transformer-based models. Extending this line of work, Borchert et al. (2023) demonstrate that textual information from company websites can be effectively leveraged using transformers to improve business failure prediction.

Recent studies have also begun to explore the potential of generative AI, such as ChatGPT, in the context of credit scoring. Wu et al. (2023) show that AI-generated text can enhance the predictive accuracy of credit models. Building on this, Yu et al. (2024) propose a GPT-LGBM framework that integrates ChatGPT with LightGBM, combining structured data with psychological features extracted from loan applicants' narratives. This approach yields substantial improvements in predictive performance. These developments highlight the growing relevance of large language models and generative AI in advancing credit risk assessment.

In light of these advances, the robustness of deep learning models has emerged as a critical concern. Despite their strong predictive performance, such models are known to be highly sensitive to adversarial attacks—small, often imperceptible perturbations to the input that can lead to substantial changes in output (Szegedy et al., 2014). While this vulnerability has been extensively studied in image classification (Goodfellow et al., 2014; Guo et al., 2017b; Nguyen et al., 2015), similar risks have been documented in natural language processing tasks (Alzantot et al., 2018; Wang et al., 2019; Zhang et al., 2020). In the textual domain, adversarial attacks typically involve subtle modifications to word choice or phrasing that preserve the original meaning for human readers but cause the model to misclassify. Techniques such as synonym substitution (Zang et al., 2019) or word-level perturbations guided by language models (Garg & Ramakrishnan, 2020) illustrate how easily text-based models can be manipulated. Empirical research confirms that such manipulations can substantially degrade model performance across various natural languages processing tasks, including sentiment classification (Xu et al., 2021) and named entity recognition (Lin et al., 2021). These vulnerabilities raise particular concerns in high-stakes applications like credit scoring, where incorrect predictions can have serious financial and ethical implications for both lenders and borrowers.

Despite the increasing adoption of deep learning and transformer-based models in credit scoring, their vulnerability to adversarial manipulation remains largely unexplored. This is particularly critical in applications involving unstructured text data, where subtle and semantically plausible changes in wording can disproportionately influence model outputs. While adversarial robustness has received growing attention in natural languages processing, little is known about how these attacks affect financial decision-making contexts, where model predictions can directly impact loan pricing, credit access, and fairness. This gap underscores the need to systematically assess the susceptibility of text-based credit scoring models to adversarial attacks and to evaluate methods—such as adversarial training—that can improve their robustness. In doing so, our study extends the literature on credit risk modeling by addressing model reliability and resilience under adversarial conditions.

3 Data

For the analysis we utilize data from Lending Club. The dataset includes loans issued between 2007 and 2014. During this time, borrowers could provide textual descriptions as part of their loan applications to describe the purpose of their loan or give a description of themselves. The full dataset, when only considering data with descriptions, consists of over 125,000 loans. The data was restricted to a subset where descriptions contained between 40 and 115 words. We set the lower limit at 40 words to ensure that each description provides meaningful context about the borrower and loan purpose. The upper limit of 115 words corresponds to the 95% percentile of the word count distribution, allowing to exclude excessively long descriptions that might introduce noise into the analysis. The final dataset contains 40,229 fully funded loans.

The dataset includes a comprehensive set of structured data, such as borrower income, credit score, and delinquencies, alongside the unstructured loan descriptions. The complete feature set aligns with commonly accepted determinants of credit defaults in peer-to-peer lending, similar to those used in prior studies by Fitzpatrick and Mues (2021) or Kriebel and Stitz (2022). Definitions and summary statistics for all structured variables in contained in the dataset are provided in Appendix A in Table 6.

In preparing the textual data for analysis, we applied standard cleaning techniques to the loan descriptions, including the removal of irrelevant metadata and converting all text to lowercase, similar to Kriebel and Stitz (2022). We choose not to apply lemmatization or stemming based on our use of BERT, a transformer-based model that captures the contextual meaning of words without requiring such preprocessing steps (Devlin et al., 2019).

The dataset exhibits a significant class imbalance with 85.5% of the loans being non-defaulted and only 14.5% being defaulted. This imbalance poses challenges for predictive modeling, as it can lead to models that are biased towards the majority class. To address this, we implement focal loss, a technique designed to mitigate bias and enhance the model's ability to accurately predict both defaulted and non-defaulted loans in subsequent analyses.

To provide an initial description of the textual data, Figure 1 displays the top 10 most frequently occurring words. These frequently used terms, such as ”credit”, ”loan”, ”pay”, and ”debt”, reflect the primary financial concerns and priorities of the borrowers. The prominence of these words suggests a strong focus on creditworthiness and financial obligations in the narratives provided by the borrowers.

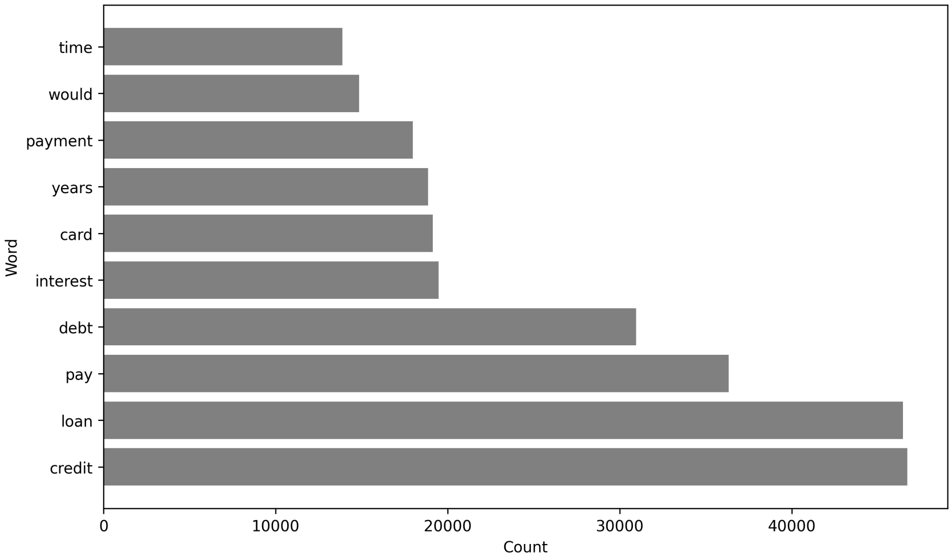


Figure 1: Bar plot showing the top 10 most frequent words in loan descriptions. Stop words have been removed to highlight more meaningful terms related to financial context.

4 Methodology

The study uses BERT as a method to exploit the textual information embedded in the loan descriptions. BERT is a common choice of transformer model that has been widely recognized for its effectiveness in capturing nuanced meanings in text. Previous studies, such as those by Stevenson et al. (2021), Kriebel and Stitz (2022), and Wu et al. (2023), have successfully utilized BERT to enhance credit default predictions by leveraging unstructured text data. We follow this path in order to study the robustness of transformer models to adversarial attacks. This section outlines further design decisions and provides a detailed explanation of the methodologies employed.

4.1 Model fine-tuning

BERT is a transformer-based model that has been pre-trained on vast amounts of text data, allowing it to understand and capture the deep contextual relationships within language (Devlin et al., 2019). The architecture of BERT is built on a stack of transformer layers, which leverage self-attention mechanisms to process each word in a sentence relative to all other words, rather than just in a sequential order. The model is pre-trained in a task-agnostic manner and subsequently fine-tuned for specific downstream applications. This approach has proven highly effective for a variety of natural language processing tasks (Devlin et al., 2019).

To adapt BERT for the task of credit default prediction, we fine-tune the model to identify high-risk loans. This involves adding a classification layer on top of BERT and retraining the model on the labeled defaulted and repaid loans. This approach allows us to retain the general language understanding BERT has learned from its pre-training, while also adapting the model to the specific nuances and patterns relevant to predicting credit default.

For the fine-tuning process, we split the dataset into three distinct subsets: a training set, a validation set, and a test set. 20,000 observations were allocated to the training set, 10,000 to the validation set, and the remaining observations were reserved for testing. The training set is used to fine-tune the BERT model, while the validation set served to monitor the model's performance and guide hyperparameter tuning, ensuring that the model generalizes well to unseen data. The test set is employed to provide an evaluation of the model's performance.

To address the challenge of class imbalance during the fine-tuning process, we employ focal loss, introduced by Lin et al. (2017). Focal loss is designed to handle class imbalance by reducing the loss contribution from well-classified instances and place more emphasis on the harder-to-classify examples. This approach is beneficial in our setting, as it mitigates the model's tendency to favor the majority class (non-defaults) and enhances its capability to correctly predict the minority class (defaults). By adding a modulating factor to the standard cross-entropy loss, focal loss down-weights the loss for correctly classified examples, allowing the model to focus more on the challenging, misclassified cases.

In our application, we follow a sample-dependent schedule for this modulating factor, as proposed by Mukhoti et al. (2020), which adjusts the model's focus based on the confidence level of each prediction. This adaptive approach helps to ensure that the model remains robust across different levels of prediction confidence, ultimately leading to a more balanced model (Mukhoti et al., 2020).

To find the hyperparameters, we perform a grid search, following the guidelines provided by Devlin et al. (2019). This grid search explores different configurations of the learning rate, the batch size, and the number of epochs to determine the combination that delivers the best performance on the validation set. The specific sets of hyperparameters considered and those selected based on the validation sample performance are detailed in Table 7 of the Appendix B.

After fine-tuning the model, we apply temperature scaling as a post-processing step to calibrate the model's output probabilities. Temperature scaling, introduced by Guo et al. (2017a), adjusts the confidence of the model's predictions by scaling the logits before applying the softmax function, which recalibrates the output probabilities. In the context of credit default prediction, it is often crucial that predicted probabilities reflect the true likelihood of default rather than just separating default from non-default. Accurate probabilities allow lenders to make informed decisions about interest rates, which is also addressed in this work. The optimal temperature parameter is determined by minimizing the cross-entropy loss on the validation set, as proposed by Mukhoti et al. (2020). The optimal temperature parameters selected based on the validation data are included in Table 7 of the Appendix B.

4.2 Ex-post word attribution analysis

To systematically identify model vulnerabilities and gain a deeper understanding of how the BERT model interprets input text, we conduct a word attribution analysis on the initial fine-tuned BERT model. This analysis highlights which tokens the model relies on most heavily when estimating default risk, thereby revealing points of susceptibility to manipulation.

We employ an integrated gradients approach (Sundararajan et al., 2017), a gradient-based attribution technique that assigns importance scores to individual input tokens. For classification tasks, integrated gradients measure how the model's prediction shifts as the input transitions from a baseline (such as zero embeddings or neutral tokens) to the actual text. By accumulating gradients along this path, the algorithm assigns an attribution score to each token, indicating how strongly that token contributes to the final prediction. Higher absolute values of these scores signify greater influence on the model's decision, with positive scores elevating the default risk estimate and negative scores reducing it (Janizek et al., 2020).

We compute these token-level attribution scores for every text sample in the training data. This process helps identify specific words or phrases that the model considers critical when forecasting a default. By examining tokens with the highest attribution scores, we gain insight into the model's primary decision drivers, uncovering both potential biases and vulnerabilities. Such knowledge is vital for understanding how malicious (or even accidental) textual changes can lead to disproportionate fluctuations in predicted default probabilities, thus underscoring the broader robustness concerns at the center of this study.

4.3 Simulating adversarial attacks and enhancing model robustness

To explore the overall robustness of the BERT-based credit default prediction model to adversarial attacks, we employ BERT-based Adversarial Examples (BAE), a method developed by Garg and Ramakrishnan (2020) to generate adversarial examples that maintain the semantic coherence and grammatical correctness of the original texts. These adversarial examples exploit model sensitivities by making subtle perturbations that are difficult for humans to detect but can mislead the model.

The BAE method starts by evaluating the importance of each token in the input text with respect to the classification decision made by the fine-tuned BERT model. This evaluation is done by measuring the decrease in the model's confidence when a token is removed, thereby identifying which tokens are most critical to the model's prediction. These critical tokens are then targeted for perturbations to maximize the impact on the model's output.

BAE applies two types of perturbations: token replacement and token insertion. In this study, we focus on token replacement. The identified critical token is masked, and the Masked Language Model (MLM) component of BERT is used to predict plausible substitutes based on the surrounding context. These candidate replacements are then filtered using semantic similarity scores derived from the Universal Sentence Encoder (USE) (Cer et al., 2018), ensuring that the modified text remains semantically close to the original. This careful filtering avoids the generation of unnatural or easily detectable adversarial inputs, addressing a key limitation of earlier methods (Alzantot et al., 2018; Di Jin et al., 2019; Ren et al., 2019).

In our setting, we apply this technique in two directions: generating adversarial samples that increase the predicted probability of default, and generating samples that decrease the predicted default probability. While a primary concern is that adversarial attacks could be used to lower default predictions and secure more favorable loan conditions, it is equally important to consider the opposite scenario. Small, unintentional changes in word choice could possibly inadvertently raise the predicted default probability, leading to such unfair lending conditions. In this sense, our analysis of adversarial text generation also reflects a broader fairness concern: that credit scoring models may disproportionately penalize some individuals based on minor linguistic nuances, unintentionally increasing their expected risk of default simply due to the phrasing of their loan applications. By evaluating both the increases and decreases in predictions, we aim to ensure that models are robust not only against intentional manipulation but also against unfair outcomes that might arise from innocent variations in text.

To mitigate the vulnerabilities revealed by this procedure, we employ adversarial training, a strategy used to enhance model resilience by exposing the model to adversarial examples during training (Kurakin et al., 2016; Morris et al., 2020). Specifically, we generate adversarial samples from the training set using the two-directional BAE approach described above. These newly created adversarial instances, some increasing and others decreasing the model's predicted probabilities of default, are then added to the original training set, producing an augmented dataset. By fine-tuning the BERT model on this augmented dataset, we aim to decrease the model's sensitivity to adversarial manipulations.

To evaluate the effectiveness of this training procedure, we generate a new set of adversarial test examples and compare the performance of the adversarially trained model to the baseline. This comparison provides insight into whether the training method successfully reduces susceptibility to adversarial influence without compromising predictive performance on clean data.

4.4 Topic modeling

To complement the use of transformer models for processing unstructured textual data, we also incorporate an alternative approach based on topic modeling. The primary idea is to condense the semantic content of loan descriptions into high-level topics, which may exhibit greater stability against minor textual perturbations. By abstracting from specific word choices to broader thematic structures, topic modeling could serve as a more resilient method for credit scoring under adversarial conditions.

For this purpose, we employ BERTopic (Grootendorst, 2022), a topic modeling technique that merges the representational power of large language models with traditional clustering and dimensionality reduction strategies. Specifically, BERTopic begins by generating dense vector embeddings of each loan description in the training and validation data obtained from our fine-tuned BERT model. These high-dimensional embeddings are then projected into a lower-dimensional space using UMAP (McInnes et al., 2018), which is designed to preserve both local and global data structure while reducing computational complexity. Next, the algorithm applies HDBSCAN (Campello et al., 2013) to detect clusters in this reduced embedding space. HDBSCAN adaptively determines the number of clusters and labels outliers as noise, making it particularly well-suited for data with variable-density regions. Finally, BERTopic represents each cluster by extracting its most distinctive keywords through class-based TF-IDF (c-TF-IDF). By comparing term frequencies across clusters rather than across individual documents, c-TF-IDF pinpoints the words that uniquely characterize each topic. Through this multi-step process BERTopic produces thematically coherent topics that serve as higher-level representations of the original loan descriptions.

To assess the predictive value of this condensed textual information, we integrate the identified topics into a logistic regression for default prediction. Specifically, we treat the topic labels as a categorical variable. By encoding each text's assigned topic, the model learns the relationship between particular topics and default outcomes without relying on the original input texts.

Because topic modeling abstracts away from specific word usage, small adversarial modifications, such as synonym replacements, are less likely to alter topic assignments. Thus, this approach may provide inherent robustness without additional adversarial defenses. Moreover, it offers improved interpretability relative to dense transformer embeddings, enabling insights into thematic drivers of credit risk. This alternative approach also enables a direct comparison of the granularity-performance trade-off: it allows us to assess whether the more granular information exploited by BERT substantially improves predictive power compared to the condensed, topic-based representations.

4.5 Aggregating structured and unstructured data

To combine the predictive power of structured and unstructured data in the credit default prediction task, we incorporate both the BERT-based predictions and the topic-level features into logistic regression models alongside the traditional structured variables. This method enables the integration of BERT-based predictions with traditional structured financial data in a modular and transparent manner. In particular, this approach reflects a potential modular inclusion where practitioners that currently still often use logistic regressions can add textual information in an additive way without entirely replacing existing systems. In order for the logistic regression to not overemphasize the predicted probabilities from the BERT models, we fit the logistic regression model using only the validation dataset.

For the analysis, we estimate four different logistic regression models: the first model uses only the structured features contained in the data; the second model incorporates both structured information and the predicted probabilities of default from the initially fine-tuned BERT model; the third model combines structured information with the predicted probabilities from the BERT model, which has been enhanced through adversarial training; the fourth model combines structured features with the categorical topic labels derived from BERTopic. Finally, temperature scaling is applied to all logistic regression models, following the procedure described in Section 4.1, to ensure the output probabilities are calibrated and interpretable.

This integration strategy enables a fair comparison across model architectures and allows us to examine how different types of textual representations contribute to credit default prediction when combined with conventional financial data.

4.6 Evaluation metrics

To evaluate the predictive performance of our models, we use two complementary metrics: the Area Under the Receiver Operating Characteristic Curve (AUC) and the Area Under the Precision-Recall Curve (AUCPR).

The AUC is a widely used measure in credit scoring and risk prediction that captures the model's ability to distinguish between defaulters and non-defaulters across all classification thresholds (Fitzpatrick & Mues, 2021; Gunnarsson et al., 2021; Kriebel & Stitz, 2022; Lessmann et al., 2015).

In addition, we report the AUCPR to provide a more informative evaluation in the presence of class imbalance. The AUCPR summarizes the trade-off between precision (positive predictive value) and recall (true positive rate) over all thresholds. This metric is especially relevant in our setting, where only 14.5% of the loans in the dataset are defaulted, leading to a strong class imbalance. Unlike AUC, the baseline for AUCPR is not fixed at 0.5 but depends on the prevalence of the positive class. In our case, the baseline AUCPR corresponds to the proportion of defaults in the dataset, i.e., 14.5%. A random classifier would thus achieve an AUCPR close to 0.145, and models must substantially exceed this baseline to demonstrate meaningful predictive power (Boyd et al., 2013; Sofaer et al., 2019).

Using both metrics allows us to detect whether adversarial effects primarily degrade the overall ranking (AUC) or disproportionately impact the identification of rare events (AUCPR). This dual-metric evaluation is particularly important for assessing robustness under adversarial conditions

Furthermore, Section 6 focuses on the economic implications of adversarial attacks by using the rates that could be offered to customers as an economic measure to assess how adversarial manipulations affect loan pricing and profitability.

5 Results

In this section, we present the empirical results of our analysis. We begin by evaluating the predictive performance of the proposed models on the original test data. We then examine the internal mechanics of the fine-tuned BERT model using word attribution analysis to identify the most influential tokens and potential vulnerabilities. Building on these insights, we assess the impact of adversarial attacks on model performance, followed by an evaluation of the effectiveness of adversarial training as a defense strategy. We further explore topic modeling as an alternative, inherently robust textual representation and analyze the thematic structure captured by the identified topics. Finally, we assess variable importance across models to quantify the contribution of structured and unstructured features to default prediction.

5.1 Performance evaluation

We first evaluate the predictive performance of the different models on clean data.

Table 1 reports both the AUC and the AUCPR for each model. AUCPR is particularly informative in our setting due to the class imbalance in the data, with only 14.5% of loans being defaulted.

The BERT model, which uses only textual loan descriptions, achieves an AUC of 0.6399 and an

AUCPR of 0.2371, indicating considerable discriminative power given that only one type of information is used. The Topic Model, based on logistic regression using topic labels as features, performs modestly below BERT with an AUC of 0.6019 and an AUCPR of 0.2288. The Structured Model, which relies exclusively on structured features contained in the data, achieves an AUC of 0.7018 and an AUCPR of 0.2806, reflecting that structured variables alone provide a solid foundation for credit default prediction.

Table 1: Model performance on the test set using AUC and AUCPR.

	BERT	**Topic**	**Structured**	**Combined**	**Combined Topic**
AUC	0.6399	0.6019	0.7018	0.7141	0.7140
AUCPR	0.2371	0.2288	0.2806	0.2914	0.2880

Combining textual and structured data yields the strongest predictive performance. The Combined Model, which integrates structured features with predictions from the BERT model, achieves an AUC of 0.7141 and an AUCPR of 0.2914, outperforming all individual models. Likewise, when topic representations are incorporated alongside structured variables in the Combined Topic Model, predictive performance improves substantially compared to the Topic Model alone, reaching an AUC of 0.7140 and an AUCPR of 0.2880.

Interestingly, while the BERT-based model outperforms the topic-based counterpart when used in isolation, the performance gap between the Combined Model and the Topic Combined Model is minimal. This suggests that despite their simplicity, topic representations can successfully capture much of the credit-relevant information embedded in the unstructured text when paired with structured data.[2]

The results confirm that incorporating textual data, either via transformer-based embeddings or topic abstractions, improves credit default prediction, particularly when combined with structured variables. This is well in line with many findings in the extant literature (Gao et al., 2018; Mai et al., 2019). Notably, the best-performing models nearly double the AUCPR relative to the random baseline of 0.145, underlining the practical value of these approaches in real-world lending settings.

[2] To formally evaluate the significance of the observed differences in predictive performance, we conduct pairwise hypothesis tests. For AUC, DeLong's test was used to assess statistical significance, while AUCPR comparisons were performed using a nonparametric bootstrap procedure with 1,000 iterations. Full test results are provided in Appendix C.

5.2 Word attributions

Building on the performance results, we now examine how individual words influence model predictions using attribution scores. This analysis sheds light on the internal logic of the BERT model and its sensitivity to linguistic variation. We illustrate this using the following two cases:

1. “This loan is to repay credit card debt that I have. I have been trying for years to pay it down, but with the high rates on the credit cards I am having problems doing it in a timely manner.”
2. “I financed significant dental work and have unexpected repairs on my home due to a water leak.”

Figure 2 and Figure 3 show the attribution scores for each word in these sentences, visualizing their contribution to the BERT model’s output. Positive attribution scores indicate that the corresponding word contributes to an increased likelihood of credit default, whereas negative scores suggest a reduced likelihood. The words with the highest absolute attribution scores are highlighted with dashed bars. These would be natural candidates for adversarial attacks.

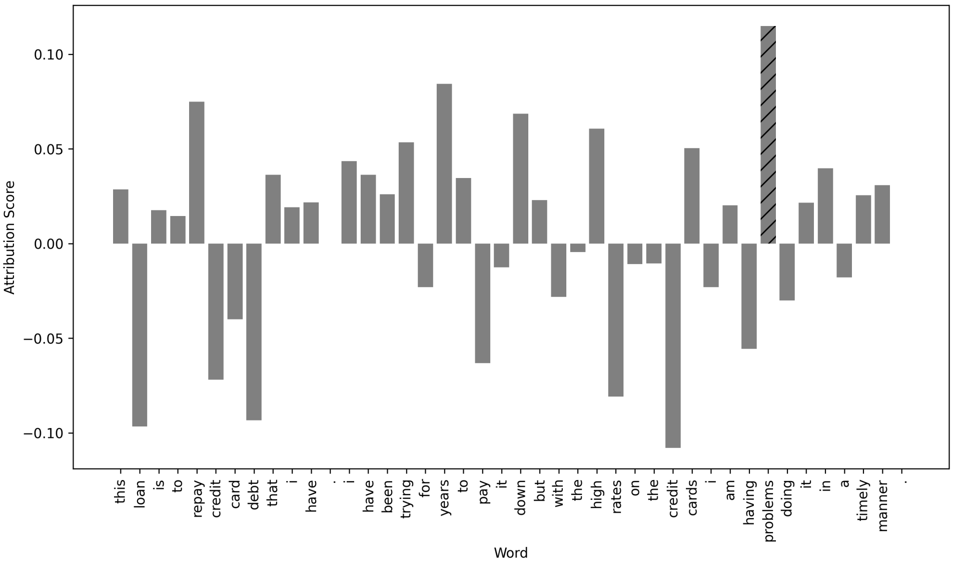


Figure 2: Bar plot displaying the word attribution scores. Words with positive attribution scores are associated with a higher likelihood of default, while words with negative scores indicate a reduced default probability.

In Figure 2, the BERT model shows a strong reliance on the term *”problems”*, which has the highest absolute attribution score of 0.1148. The model associates this term with financial distress and a higher likelihood of default. In contrast, other terms such as *”loan”*, *”debt”*, and *”credit”* have large negative attribution scores, indicating that these words reduce the default probability in this context. This could suggest that the model learned to associate references to financial planning with a higher creditworthiness. The combination of positive and negative attributions across different terms reflects the nuanced way in which the model integrates contextual information to form its predictions.

In Figure 3, the word *”financed”* exhibits a strong negative attribution score of -0.1266, indicating that the model associates it with more responsible financial behavior, thereby lowering the predicted probability of default. Similarly, the words *”repairs”* and *”home”* also have negative attribution scores, possibly reflecting an association with financial responsibility or potentially collateral, though to a lower extent compared to *”financed”*. In contrast, the word *”unexpected”* contributes positively to the default prediction, highlighting a potential risk factor as it could imply an unforeseen financial burden. This pattern demonstrates the model's sensitivity to contextual cues, where terms related to planned expenditures or property might reduce the default probability, while expressions suggesting unpredictability increase the perceived credit risk.

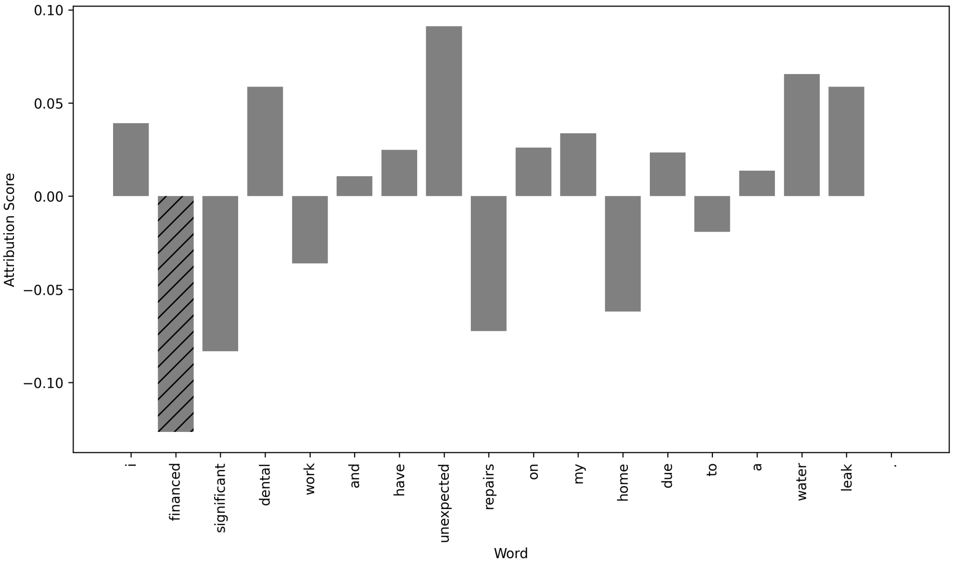


Figure 3: Bar plot displaying word attribution scores. Words with positive attribution scores are associated with a higher likelihood of default, while words with negative scores indicate reduced default probability.

To further analyze the model's sensitivity to adversarial attacks, we investigate how substituting the most influential key words with synonyms impacts the associated attribution scores and the predicted probabilities. For Figure 2, replacing the term *“problems”* with the synonym *“trouble”* would result in a notable drop in the attribution score from 0.1148 to 0.0303, leading to a substantial decrease in the default probability from 21.69% to 14.08%. Similarly, using *"difficulties"* instead of *"problems"* would yield an attribution score of 0.0713 and a default probability of 16.82%. Despite the seeming similarity between these terms, the model's outputs shift dramatically, indicating a strong dependence on specific high-impact words. A potential borrower that would like to obtain a favorable loan rate and has the technical skills to detect this vulnerability could thus use this for the own benefit.

A similar pattern but with another direction emerges in Figure 3. When the term *"financed"* is replaced with *"funded"*, the attribution score would change from -0.1266 to -0.0905, increasing the default probability from 22.93% to 26.22%. Using the phrase *"paid for"* would lead to an attribution score of -0.1076 and a corresponding default probability of 24.03%. These changes also highlight the model's sensitivity to subtle linguistic nuances and suggest that seemingly minor variations can alter its interpretation of the borrower's creditworthiness. From a borrower perspective, replacing *"problems"* with one of these synonyms without being aware of the model associations could lead to receiving a considerably worse loan offer potentially disadvantaging customers.

Addressing these vulnerabilities through using more robust models thus seems important to assess for potentially enhancing model resilience and ensuring more reliable and fair loan pricing.

5.3 Exposure to adversarial attacks

Next, the results under the adversarial conditions are presented in Table 2. This reflects the situation where a lender would make decisions based on a model that uses unstructured text data and faces potential adversarial attacks. The table presents both AUC and AUCPR values for each models on both the original and manipulated samples.

Table 2: Model performance under original and adversarial conditions using AUC and AUCPR. Significance levels for AUC differences are calculated using DeLong's test; AUCPR differences use nonparametric bootstrapping. They are denoted as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	AUC		**AUCPR**		**Difference**	
Model	Original	Adversarial	Original	Adversarial	AUC	AUCPR
BERT	0.6399	0.5926	0.2371	0.1940	0.0473***	0.0431***
Topic	0.6019	0.5988	0.2288	0.2183	0.0031*	0.0105***
Combined	0.7141	0.6954	0.2914	0.2756	0.0187***	0.0158***
Topic Combined	0.7140	0.7123	0.2880	0.2818	0.0017**	0.0062**

The BERT model shows the steepest decline, with an AUC drop of 0.0473 and an AUCPR decrease of 0.0431, reflecting high sensitivity to minor semantic perturbations. In contrast, the Topic Model exhibits only minimal performance degradation, particularly in AUC, suggesting that topic abstractions are less sensitive to adversarial attacks.

Interestingly, the Combined Model, despite its superior performance on clean data, falls below the Structured Model in AUC and AUCPR under adversarial conditions (0.6954 vs. 0.7018). This implies that adversarial manipulations targeting the textual component can offset the gains achieved by integrating unstructured data. In contrast, the Topic Combined Model not only retains strong predictive performance but also continues to outperform the Structured Model even when exposed to adversarial inputs. This highlights the comparative robustness of topic-based representations when used in combination with structured features.

These results suggest that while the integration of textual data can enhance prediction accuracy, it may also introduce vulnerabilities depending on how that information is encoded. Topic-based abstractions appear to offer a more stable alternative to fine-grained embeddings in adversarial environments, likely due to their reduced sensitivity to specific word choices.

To further illustrate the impact of adversarial attacks on the model predictions, Figure 4 visualizes the changes in predicted default probabilities under adversarial attacks for the Combined Model and the Topic Combined Model. Each plot shows the distributions of predicted probabilities for the original data, manipulated samples designed to increase the predicted default probabilities, and manipulated samples aimed at decreasing the default probabilities.

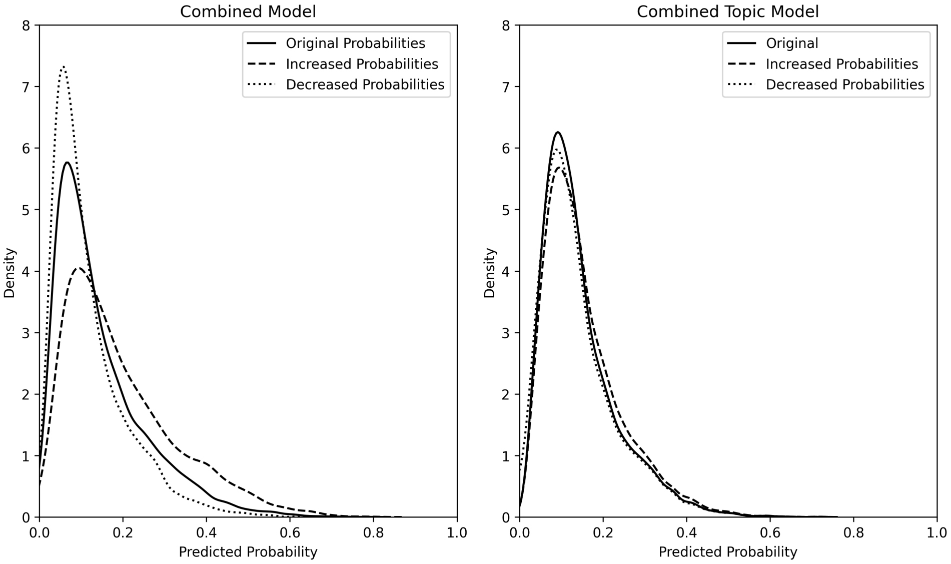


Figure 4: Kernel density estimates of predicted probabilities for the Combined Model (left) and the Topic Combined Model (right) under original and adversarial conditions.

For the Combined Model, adversarial manipulations produce visible shifts in the predicted probability distribution. Attacks intended to raise default probabilities result in a distribution with a location that is moved more towards higher probabilities and with a lower right-skewness, indicating heightened risk perception. Conversely, adversarial conditions to lower default probabilities shift the location of the distribution toward lower predicted values and increase the right-skewness, potentially leading to underestimated risk. These shifts demonstrate the vulnerability of models combining structured and unstructured data to adversarial attacks, which could significantly alter the perception of borrower creditworthiness.

In contrast, the Topic Combined Model demonstrates much greater stability across conditions. The distributions for the manipulated samples remain close to the original, indicating that topic representations filter out much of the sensitivity to lexical perturbations. This robustness, combined with strong baseline performance, positions the topic-based approach as a viable and interpretable alternative to transformer-based representations in adversarial settings.

5.4 Performance of the robust model

To assess the effectiveness of adversarial training in improving model robustness, we evaluate both the robust BERT model and the robust Combined Model under clean and adversarial conditions. Table 3 presents AUC and AUCPR values for both models.

Table 3: Performance of robust models under original and adversarial conditions.

	AUC		**AUCPR**		**Difference**	
Model	Original	Adversarial	Original	Adversarial	AUC	AUCPR
Robust BERT	0.6363	0.6129	0.2332	0.2111	0.0234***	0.0221***
Robust Combined	0.7150	0.7086	0.2921	0.2860	0.0064	0.0061

Both robust models exhibit smaller performance declines under adversarial attacks than their non-robust counterparts. The robust Combined Model, in particular, shows a notably reduced AUC drop (0.0064) and AUCPR drop (0.0061), indicating improved stability. Importantly, it continues to outperform the Structured Model under adversarial conditions in both AUC (0.7086 vs. 0.7018) and AUCPR (0.2860 vs. 0.2806), underscoring the benefit of incorporating unstructured information in a robust way.

When compared to the Topic Combined model, the robust Combined Model performs slightly better on the original data in both AUC (0.7150 vs. 0.7140) and AUCPR (0.2921 vs. 0.2880).

However, under adversarial conditions, the Topic Combined model achieves a marginally higher AUC (0.7123 vs. 0.7086), while the robust Combined Model maintains a slightly higher AUCPR (0.2860 vs. 0.2818). These differences suggest a trade-off between the granularity of transformer-based embeddings and the inherent stability of topic-based representations. Adversarial training enables the BERT-based model to reach robustness levels comparable to the topic-based approach, while preserving its superior predictive performance on clean data.

To complement the numerical findings, Figure 5 displays kernel density estimates of predicted default probabilities under original and adversarial conditions for the Combined Model, the Robust Combined Model, and the Topic Combined Model. Each subplot illustrates the distribution of predicted probabilities for the original data, as well as adversarially manipulated inputs designed to either increase or decrease the predicted risk of default.

The distributions for the Combined Model (left panel) show substantial shifts in both directions, indicating significant vulnerability to adversarial perturbations. In contrast, the Robust Combined Model (center panel) exhibits notably more stable distributions, with smaller shifts in both increased and decreased scenarios. This supports the conclusion that adversarial training effectively mitigates model sensitivity to semantic manipulations.

Interestingly, the Topic Combined Model (right panel) shows similarly stable behavior under adversarial conditions, with distributional shifts that are even more contained than those of the robust BERT-based model. This visual evidence complements the performance metrics, suggesting that topic-based representations offer inherent robustness to adversarial inputs.

Overall, these results reinforce the dual pathways to adversarial resilience: training neural models with adversarial objectives, or using higher-level, interpretable representations that are less sensitive to lexical variation.

5.5 Topic interpretation

To gain deeper insights into the thematic structure captured by the topic modeling approach, we analyze the topics learned from the data. Table 4 summarizes the eleven identified topics, including their manually assigned labels, the five most representative words per topic, and the number of instances assigned to each.

The topic modeling approach reveals that borrowers' narratives can be meaningfully grouped into

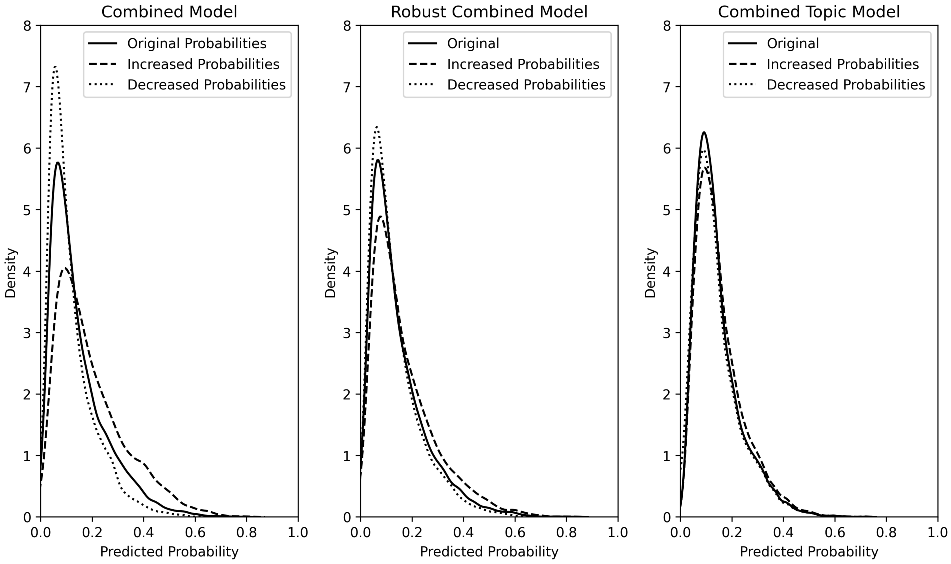


Figure 5: Kernel density estimates of predicted probabilities for the Combined Model (left), the Robust Combined Model (center), and the Topic Combined Model (right) under original and adversarial conditions.

Table 4: Summary of the topics derived from borrower loan descriptions. Topic labels are manually assigned based on the five most representative words.

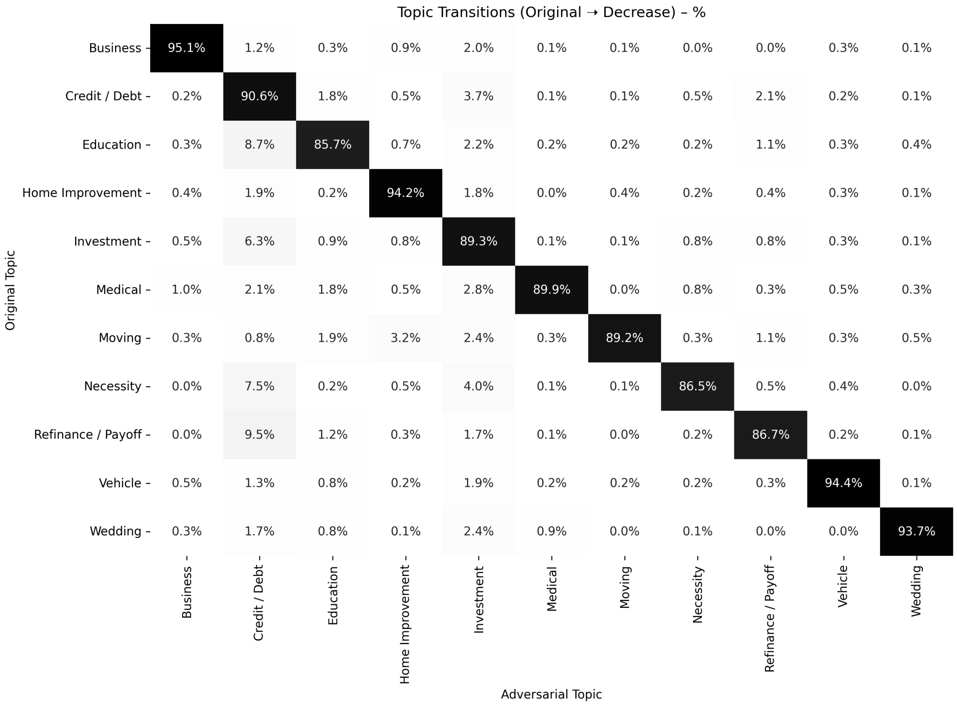
Topic Label	**Top Words**	**Count**
Credit / Debt	`credit, loan, consolidate, debt, pay`	15,559
Refinance / Payoff	`rate, refinance, payoff, apr, reduce`	3,487
Investment	`investor, invest, fund, start, credit`	3,290
Education	`university, graduate, degree, education, pay`	2,112
Home Improvement	`house, need, property, repair, build`	1,670
Business	`business, market, product, expand, service`	1,112
Vehicle	`vehicle, car, finance, bike, honda`	965
Necessity	`need, pay, urgent, help, repair`	664
Wedding	`ring, engagement, loan, wedding, proposal`	576
Medical	`surgery, medical, health, hospital, expense`	283
Moving	`apartment, relocation, rental, expense, relocate`	282

interpretable themes. Dominant categories include Credit / Debt, Refinance / Payoff, and Investment, which together represent a substantial portion of the dataset. Other topics, such as Education, Home Improvement, Medical, and Moving, reflect more specific financial needs. Notably, the topics align well with real-world use cases and borrower intents observed in consumer lending.

To assess the robustness of these topic representations, we examine how topic assignments change under adversarial manipulation-specifically, when the text is perturbed to reduce the predicted probability of default.[3] Figure 6 shows the percentage of samples that transition from their original topic

[3] Results for adversarial samples aimed at increasing predicted probabilities are qualitatively similar and are provided in Appendix D.

to a new topic under such manipulations.



	Business	Credit / Debt	Education	Home Improvement	Investment	Medical	Moving	Necessity	Refinance / Payoff	Vehicle	Wedding
Business	95.1%	1.2%	0.3%	0.9%	2.0%	0.1%	0.1%	0.0%	0.0%	0.3%	0.1%
Credit / Debt	0.2%	90.6%	1.8%	0.5%	3.7%	0.1%	0.1%	0.5%	2.1%	0.2%	0.1%
Education	0.3%	8.7%	85.7%	0.7%	2.2%	0.2%	0.2%	0.2%	1.1%	0.3%	0.4%
Home Improvement	0.4%	1.9%	0.2%	94.2%	1.8%	0.0%	0.4%	0.2%	0.4%	0.3%	0.1%
Investment	0.5%	6.3%	0.9%	0.8%	89.3%	0.1%	0.1%	0.8%	0.8%	0.3%	0.1%
Medical	1.0%	2.1%	1.8%	0.5%	2.8%	89.9%	0.0%	0.8%	0.3%	0.5%	0.3%
Moving	0.3%	0.8%	1.9%	3.2%	2.4%	0.3%	89.2%	0.3%	1.1%	0.3%	0.5%
Necessity	0.0%	7.5%	0.2%	0.5%	4.0%	0.1%	0.1%	86.5%	0.5%	0.4%	0.0%
Refinance / Payoff	0.0%	9.5%	1.2%	0.3%	1.7%	0.1%	0.0%	0.2%	86.7%	0.2%	0.1%
Vehicle	0.5%	1.3%	0.8%	0.2%	1.9%	0.2%	0.2%	0.2%	0.3%	94.4%	0.1%
Wedding	0.3%	1.7%	0.8%	0.1%	2.4%	0.9%	0.0%	0.1%	0.0%	0.0%	93.7%

Figure 6: Transition matrix of topic assignments under adversarial perturbations aimed at decreasing predicted default probabilities. Each row represents the original topic; each column the reassigned topic. Diagonal values indicate the proportion of samples that retained their original topic.

Overall, the results demonstrate a high degree of topic stability. For all topics, more than 85% of samples retain their original assignment even after adversarial manipulation. For instance, 94.4% of ”Vehicle” loans, 94.2% of ”Home Improvement” loans, and 93.7% of ”Wedding”-related texts remain in their original categories. Even more general topics, such as ”Credit / Debt” and ”Education,” show strong retention rates of over 85%. The relatively small off-diagonal values suggest that adversarial perturbations tend to preserve the broader semantic structure of the text, which likely contributes to the robustness observed in the Topic Combined model’s predictive performance.

These findings underscore the advantage of topic modeling in adversarial contexts: by abstracting away from fine-grained lexical patterns, topic representations offer both interpretability and inherent resilience to manipulation. This stability further enhances their appeal for high-stakes decision-making environments, such as consumer credit scoring.

5.6 Variable importance

To further examine how textual features contribute to credit risk prediction in the presence of structured data, we analyze variable importance scores derived from a permutation-based approach (Breiman, 2001; Fisher et al., 2019).[4] Figure 7 presents the permutation importance of each feature in the Combined Model, the Robust Combined Model, and the Combined Topic Model.

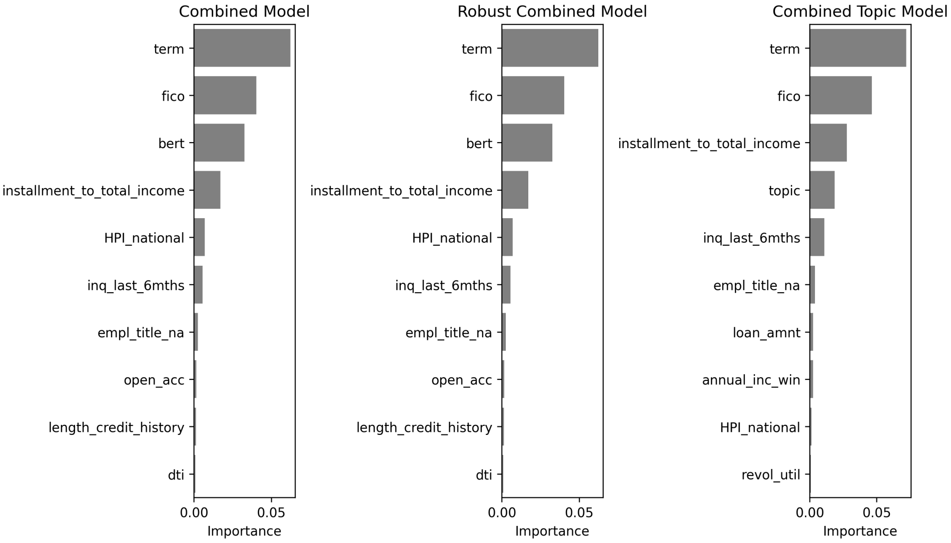


Figure 7: Permutation-based variable importance for the Combined Model (left), Robust Combined Model (center), and Topic Combined Model (right). Higher values indicate greater contribution to model performance.

Across all three models, *term* and *fico* emerge as the most influential features, which aligns with domain knowledge regarding loan length and borrower credit score. In both the Combined Model and the Robust Combined Model, the *bert* feature—representing the BERT-based textual prediction—ranks as the third most important variable. This confirms that unstructured text information provides substantial predictive value even in the presence of rich structured data.

In the Topic Combined Model, the *topic* variable—representing the set of topic-based dummy variables—also contributes meaningfully to the prediction task, though its importance is slightly lower than that of the BERT feature in the Combined Models. This suggests that while BERT embeddings capture a more granular signal from text, the topic-based representations still encode substantial credit-relevant information.

Notably, the importance scores for structured variables remain relatively consistent across models,

[4] We opt for permutation-based variable importance rather than more classical measures (e.g., absolute values of estimated coefficients) because permutation-based methods provide a single, aggregated importance score for each predictor (including multi-level factors), whereas coefficient-based methods yield multiple scores for a single categorical variable.

indicating that the inclusion of unstructured features does not obscure or destabilize the contribution of traditional credit attributes.

Overall, these results highlight that both BERT-derived and topic-based textual features significantly enhance the predictive signal in credit scoring models. The permutation-based analysis offers additional transparency into how different information sources interact, reinforcing the practical value of integrating text into credit risk assessments.

6 Economic Impact of Adversarial Attacks on Loan Pricing

In this section, we evaluate the impact of adversarial attacks on loan pricing, focusing on the financial implications for both lenders and borrowers. From a lender's perspective, the offered interest rate should reflect the level of risk in the projected cash flows. Adversarial attacks that manipulate predicted default probabilities can skew these estimates, leading to either over- or underestimation of a loan's profitability. Overestimated risks may result in lenders offering higher interest rates, while underestimated risks could cause loans to be offered at an inaccurately low rate.

For each loan, we determine the appropriate interest rate based on the predicted default probabilities and the corresponding installments. For the loss given default, we use the standard value of 0.45 from the Basel framework. The cash flows are then discounted at spot interest rates tied to the loan's issue date. The appropriate interest rate is obtained by finding the rate that equates the present value of these cash flows to the initial loan amount. Since lenders will further additionally require a slightly higher interest including a profit margin, we assume a profit margin of 6% that is added to the interest rates. We conduct this analysis on both the clean test samples and the manipulated samples, separated by manipulations aimed at increasing or decreasing predicted default risk.

Table 5: Mean appropriate interest rates for the Combined, Robust Combined, and Combined Topic models under clean and adversarial conditions. The table also reports the average changes in interest rates resulting from adversarial manipulations that either increase or decrease predicted default probabilities. Significance levels for the differences are calculated using paired t-tests and adjusted using Bonferroni correction for multiple testing. They are denoted as * $p_{adj} < 0.1$, ** $p_{adj} < 0.05$, *** $p_{adj} < 0.01$.

	IR	**IR decr.**	**IR incr.**	**Diff decr.**	**Diff incr.**
Combined Model	10.07%	9.11%	11.85%	-0.96***	1.78***
Robust Combined Model	10.38%	10.00%	11.41%	-0.38***	1.03***
Combined Topic Model	10.16%	10.04%	10.41%	-0.12***	0.25***

Table 5 shows that adversarial attacks can significantly impact loan profitability. The Combined Model exhibits the largest shifts: an average underpricing of 0.96 percentage points when default probabilities are artificially lowered, and an overpricing of 1.78 percentage points when they are increased. Such mispricings can distort credit allocation, introduce lender risk, and generate fairness concerns for borrowers.

By contrast, the Robust Combined Model, enhanced via adversarial training, significantly dampens these distortions. Interest rate shifts are nearly halved, with changes of only -0.38 and +1.03 percentage points respectively. This demonstrates that adversarial training not only enhances classification robustness, but also meaningfully mitigates economic misjudgments.

Interestingly, the Combined Topic Model shows the smallest pricing shifts: only -0.12 and +0.25 percentage points on average. While its predictive performance is marginally lower than the other combined models on clean data, the Combined Topic Model exhibits a markedly higher degree of robustness under adversarial conditions.

From the borrower's perspective, these pricing shifts matter. In the worst-case scenario, a minor change in wording could unjustly elevate a borrower's predicted default risk, resulting in loan rejection or an inflated interest rate. Conversely, strategic language manipulation could lead to artificially lowered rates, undermining model fairness and lender profitability. Both cases raise concerns about transparency, reliability, and equitable credit access.

In summary, adversarial attacks can induce substantial economic distortions in loan pricing, particularly in models that rely heavily on unstructured textual data. Adversarial training improves resilience, minimizing the potential economic misjudgement. Topic-based models, while simpler, offer compelling robustness and interpretability benefits. These results underscore the need for robustness not only as a technical property, but as a central pillar of responsible, economically sound credit modeling.

7 Conclusion

This study contributes to the body of research on credit default prediction by addressing a critical gap: the robustness of machine learning models, particularly transformer-based models, that incorporate unstructured text data. While the inclusion of textual borrower narratives has been shown to significantly improve predictive performance, our findings reveal a trade-off between predictive gains and increased vulnerability to adversarial manipulation. Even semantically subtle changes—imperceptible to human reviewers—can drastically distort model outputs, thereby challenging the reliability of these systems in real-world financial decision-making.

Our findings demonstrate that transformer-based models leveraging unstructured data are highly sensitive to subtle semantic changes. Using adversarial text generation techniques, we show that even minor, human-imperceptible modifications can substantially shift default predictions. This vulnerability poses a significant risk in practical lending environments, where textual inputs may be altered deliberately or simply vary due to differences in phrasing, writing style, or linguistic background. To better understand these sensitivities, we apply explainable artificial intelligence methods and find that model outputs are disproportionately driven by a small number of influential tokens. This raises broader concerns around model fairness and consistency: borrowers with otherwise similar profiles may receive divergent credit assessments based on small, semantically equivalent differences in expression.

We further demonstrate that these vulnerabilities can be effectively mitigated through adversarial training. By exposing models to manipulated examples during training, we significantly improve their robustness and reducing performance degradation under adversarial conditions. As an alternative approach, topic modeling provides a more abstract representation of text that, while less granular, offers greater interpretability and inherent stability against adversarial perturbations.

Finally, we evaluate the economic implications of adversarial attacks. Our analysis shows that manipulated inputs can distort loan pricing, leading to financial losses for lenders and potentially worse conditions or unfair rejections for borrowers. These distortions are markedly reduced in robust model variants, highlighting the practical importance of resilience not just for model performance, but for equitable and reliable credit allocation. From a broader perspective, our results speak to a core modeling challenge in operations research: how to integrate unstructured data in ways that are both powerful and robust. The comparison between fine-grained transformer models and topic-based abstractions illustrates the trade-off between performance, interpretability, and stability. As the adoption of unstructured data accelerates across domains, the need for robust, transparent, and fair models becomes paramount.

In sum, we advocate for credit scoring systems that are designed with robustness as a first-order objective, not just to safeguard against manipulation, but to ensure reliable, equitable, and trustworthy financial decision-making in increasingly data-rich environments.

References

Agarwal, S., Chen, V. Y. S., & Zhang, W. (2016). The information value of credit rating action reports: A textual analysis. *Management Science*, *62*(8), 2218–2240. https://doi.org/10.1287/mnsc.2015.2243

Ahmadi, Z., Martens, P., Koch, C., Gottron, T., & Kramer, S. (2018). Towards bankruptcy prediction: Deep sentiment mining to detect financial distress from business management reports. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 293–302. https://doi.org/10.1109/DSAA.2018.00040

Alzantot, M., Sharma, Y., Elgohary, A., Ho, B.-J., Srivastava, M., & Chang, K.-W. (2018). Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*. https://doi.org/10.48550/arXiv.1804.07998

Baesens, B., van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. *Machine learning*, *81*, 121–148. https://doi.org/10.1007/s10994-010-5188-5

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies*, *33*(7), 2845–2897. https://doi.org/10.1093/rfs/hhz099

Borchert, P., Coussement, K., De Caigny, A., & De Weerdt, J. (2023). Extending business failure prediction models with textual website content using deep learning. *European Journal of Operational Research*, *306*(1), 348–357.

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, 451–466.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Pacific-Asia conference on knowledge discovery and data mining*, 160–172.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder. https://arxiv.org/abs/1803.11175

Chen, X., Huang, B., & Ye, D. (2018). The role of punctuation in P2P lending: Evidence from China. *Economic Modelling*, *68*, 634–643. https://doi.org/10.1016/j.econmod.2017.05.007

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447–1465. https://doi.org/10.1016/j.ejor.2006.09.100

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Di Jin, Jin, Z., Zhou, J. T., & Szolovits, P. (2019). Is BERT really robust? Natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*, *2*(10). https://doi.org/10.48550/arXiv.1907.11932

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & Kammler, J. (2016). Description-text related soft information in peer-to-peer lending–evidence from two leading European platforms. *Journal of Banking & Finance*, *64*, 169–187. https://doi.org/10.1016/j.jbankfin.2015.11.009

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. *European Journal of Operational Research*, *249*(2), 427–439. https://doi.org/10.1016/j.ejor.2015.09.014

Fitzpatrick, T., & Mues, C. (2021). How can lenders prosper? Comparing machine learning approaches to identify profitable peer-to-peer loan investments. *European Journal of Operational Research*, *294*(2), 711–722. https://doi.org/10.1016/j.ejor.2021.01.047

Gao, Q., Lin, M., & Sias, R. W. (2018). Words matter: The role of texts in online credit markets. *Journal of Financial and Quantitative Analysis*. https://doi.org/10.1017/S0022109022000850

Garg, S., & Ramakrishnan, G. (2020). BAE: BERT-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*. https://doi.org/10.48550/arXiv.2004.01970

Goodfellow, I., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. https://doi.org/10.48550/arXiv.1412.6572

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Gunnarsson, B. R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, *295*(1), 292–305. https://doi.org/10.1016/j.ejor.2021.03.006

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017a). On calibration of modern neural networks. *International Conference on Machine Learning*, 1321–1330. https://doi.org/10.48550/arXiv.1706.04599

Guo, C., Rana, M., Cisse, M., & Van Der Maaten, L. (2017b). Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*. https://doi.org/10.48550/arXiv.1711.00117

Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell me a good story and I may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research*, *48*(SPL), 138–149. https://doi.org/10.1509/jmkr.48.SPL.S138

Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science*, *62*(6), 1554–1577. https://doi.org/10.1287/mnsc.2015.2181

Janizek, J. D., Sturmfels, P., & Lee, S.-I. (2020). Explaining explanations: Axiomatic feature interactions for deep networks. https://doi.org/10.48550/arXiv.2002.04138

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, *266*(1), 511–529. https://doi.org/10.1007/s10479-017-2668-z

Korangi, K., Mues, C., & Bravo, C. (2023). A transformer-based model for default prediction in mid-cap corporate markets. *European Journal of Operational Research*, *308*(1), 306–320.

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, *297*(3), 1083–1094.

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, *302*(1), 309–323. https://doi.org/10.1016/j.ejor.2021.12.024

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–a review. *European Journal of Operational Research*, *180*(1), 1–28. https://doi.org/10.1016/j.ejor.2006.08.043

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*. https://doi.org/10.48550/arXiv.1611.01236

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lin, B. Y., Gao, W., Yan, J., Moreno, R., & Ren, X. (2021). Rockner: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. *arXiv preprint arXiv:2109.05620*. https://doi.org/10.48550/arXiv.2109.05620

Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, *59*(1), 17–35. https://doi.org/10.1287/mnsc.1120.1560

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, 2980–2988. https://doi.org/10.48550/arXiv.1708.02002

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, *274*(2), 743–758. https://doi.org/10.1016/j.ejor.2018.10.024

Matin, R., Hansen, C., Hansen, C., & Mølgaard, P. (2019). Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications*, *132*, 199–208. https://doi.org/10.1016/j.eswa.2019.04.071

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426.*

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Di Jin, & Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. *arXiv preprint arXiv:2005.05909.* https://doi.org/10.48550/arXiv.2005.05909

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, *33*, 15288–15299. https://doi.org/10.48550/arXiv.2002.09437

Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research*, *56*(6), 960–980. https://doi.org/10.1177/0022243719852959

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 427–436.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, *74*, 26–39. https://doi.org/10.1016/j.asoc.2018.10.004

Ren, S., Deng, Y., He, K., & Che, W. (2019). Generating natural language adversarial examples through probability weighted word saliency. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1085–1097. https://doi.org/10.18653/v1/P19-1103

Shi, Y., Qu, Y., Chen, Z., Mi, Y., & Wang, Y. (2024). Improved credit risk prediction based on an integrated graph representation learning approach with graph transformation. *European Journal of Operational Research*, *315*(2), 786–801.

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577.

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, *295*(2), 758–771. https://doi.org/10.1016/j.ejor.2021.03.008

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319–3328. https://dl.acm.org/doi/10.5555/3305890.3306024

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. https://arxiv.org/abs/1312.6199

Tsai, M.-F., & Wang, C.-J. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, *257*(1), 243–250. https://doi.org/10.1016/j.ejor.2016.06.069

Wang, W., Wang, R., Wang, L., Wang, Z., & Ye, A. (2019). Towards a robust deep neural network in texts: A survey. *arXiv preprint arXiv:1902.07285*. https://doi.org/10.48550/arXiv.1902.07285

Wang, Z., Jiang, C., Zhao, H., & Ding, Y. (2020). Mining semantic soft factors for credit risk evaluation in peer-to-peer lending. *Journal of Management Information Systems*, *37*(1), 282–308. https://doi.org/10.1080/07421222.2019.1705513

Wu, Z., Dong, Y., Li, Y., & Shi, B. (2023). Unleashing the power of text for credit default prediction: Comparing human-generated and ai-generated texts. *Available at SSRN 4601317*. https://dx.doi.org/10.2139/ssrn.4601317

Xia, Y., He, L., Li, Y., Liu, N., & Ding, Y. (2020). Predicting loan default in peer-to-peer lending using narrative data. *Journal of Forecasting*, *39*(2), 260–280. https://doi.org/10.1002/for.2625

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Xu, Y., Zhong, X., Yepes, A. J., & Lau, J. H. (2021). Grey-box adversarial attack and defence for sentiment classification. *arXiv preprint arXiv:2103.11576*. https://doi.org/10.48550/arXiv.2103.11576

Yu, L., Bai, X., & Chen, Z. (2024). Gpt-Lgbm: A Chatgpt-based integrated framework for credit scoring with textual and structured data. https://dx.doi.org/10.2139/ssrn.4671511

Zandi, S., Korangi, K., Óskarsdóttir, M., Mues, C., & Bravo, C. (2025). Attention-based dynamic multilayer graph neural networks for loan default prediction. *European Journal of Operational Research*, *321*(2), 586–599.

Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., & Sun, M. (2019). Word-level textual adversarial attacking as combinatorial optimization. *arXiv preprint arXiv:1910.12196*. https://doi.org/10.48550/arXiv.1910.12196

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3), 1–41. https://doi.org/10.1145/3374217

A Structured Features

Table 6: Summary statistics for continuous and binary variables. The reported metrics include the arithmetic mean (mean), standard deviation (sd), minimum (min), and maximum (max) values. The total number of observations (N) is 40,229. The annual earnings are winsorized at the 1st and 99th percentiles.

Attribute	**Unit**	**Mean**	**SD**	**Min**	**Max**
Open accounts	Count of accounts	10.449	4.633	0	53
Debt ratio	Payment to income ratio	15.72	7.42	0.00	36.82
Loan amount	Amount in USD	13,883.42	7,847.76	700.00	35,000.00
Employment title reported	Binary indicator	0.051	0.220	0	1
Verified income	Binary indicator	0.629	0.483	0	1
FICO score	Interval center	707.523	33.653	632	848
Revolving credit utilization	Credit used/available	50.07	28.17	0.00	113.00
Delinquency record	Count of delinquencies	0.19	0.61	0.00	18.00
Loan term	Time in months	41.629	10.169	36	60
Credit history length	Time in month	171.504	79.646	36	684
Annual earnings	USD (winsorized)	69,500.280	36,826.810	18,000	230,000
Revolving debt balance	Amount in USD	15,400.790	19,328.970	0	1,746,716
Recent credit inquiries	Number of inquiries (6 months)	0.819	1.045	0	8
State unemployment	Unemployment rate	8.626	1.882	2.400	14.000
Housing price index	Percent change (year-over-year)	-5.325	10.838	-26.790	17.870
Income to installment ratio	Fraction of income	0.007	0.003	0.0001	0.027
Job tenure unknown	Binary indicator	0.023	0.150	0	1
Public record defaults	Count of defaults	0.071	0.314	0	17
Loan default status	Binary indicator	0.145	0.352	0	1

B Hyperparameters

Table 7: Hyperparameter search spaces. This table shows the considered hyperparameter spaces for the methods used in this study. The final hyperparameter combination is chosen based on validation sample performance. The best hyperparameter value is shown in the fourth column.

Model	**Hyperparameter**	**Parameter Space**	**Best Hyperparameter**
BERT	Batch size	16, 32	0.5
	Learning rate	5e-5, 3e-5, 2e-5	2e-5
	Epochs	2, 3, 4	4
	Temperature	-	1.7229
Robust BERT	Batch size	16, 32	0.4
	Learning rate	5e-5, 3e-5, 2e-5	3e-5
	Epochs	2, 3, 4	2
	Temperature	-	1.8193
Model	**Hyperparameter**	**Parameter space**	**Best hyperparameter**
Structured Model	Temperature	-	1.8741
Combined Model	Temperature	-	2.0979
Robust Combined Model	Temperature	-	2.0318

C Statistical Significance Tests for Model Performance

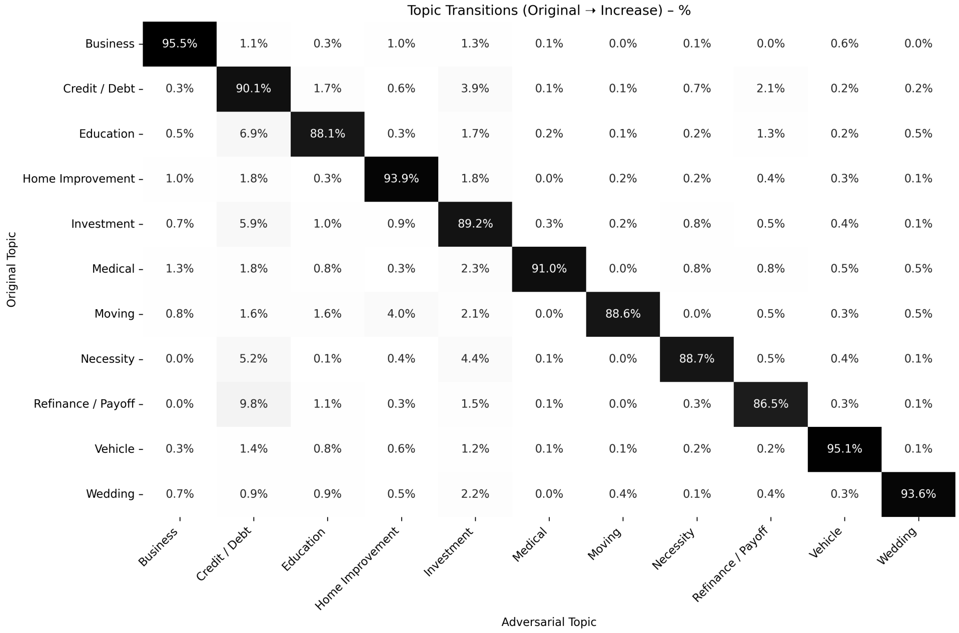
Model A	**Model B**				
	BERT	Topic	Structured	Combined	Combined Topic
BERT	-	0.0380***	-0.0619	-0.0742	-0.0741
Topic	-0.038	-	-0.0999	-0.1122	-0.1121
Structured	0.0619***	0.0999***	-	-0.0123	-0.0122
Combined	0.0742***	0.1122***	0.0123***	-	0.0001
Topic Combined	0.0741***	0.1121***	0.0122***	-0.0001	-

Table 8: Pairwise AUC differences between models based on the test set. Each cell reports the difference $\text{AUC}_{\text{Model A}} - \text{AUC}_{\text{Model B}}$. Statistical significance is assessed using one-sided DeLong tests. Reported p-values are adjusted for multiple comparisons using Bonferroni correction. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Model A	**Model B**				
	BERT	Topic	Structured	Combined	Combined Topic
BERT	-	0.0083	-0.0435	-0.0543	-0.0509
Topic	-0.0083	-	-0.0518	-0.0626	-0.0592
Structured	0.0435***	0.0518***	-	-0.0108	-0.0074
Combined	0.0543***	0.0626***	0.0108	-	0.0034
Topic Combined	0.0509***	0.0592***	0.0074	-0.0034	-

Table 9: Pairwise AUCPR differences between models based on the test set. Each cell reports the difference $\text{AUCPR}_{\text{Model A}} - \text{AUCPR}_{\text{Model B}}$. Statistical significance is assessed using nonparametric bootstrapping with 1,000 resamples. Reported p-values are adjusted for multiple comparisons using Bonferroni correction. Significance levels are denoted as: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

D Topic Transition Matrix – Increased Probabilities



	Business	Credit / Debt	Education	Home Improvement	Investment	Medical	Moving	Necessity	Refinance / Payoff	Vehicle	Wedding
Business	95.5%	1.1%	0.3%	1.0%	1.3%	0.1%	0.0%	0.1%	0.0%	0.6%	0.0%
Credit / Debt	0.3%	90.1%	1.7%	0.6%	3.9%	0.1%	0.1%	0.7%	2.1%	0.2%	0.2%
Education	0.5%	6.9%	88.1%	0.3%	1.7%	0.2%	0.1%	0.2%	1.3%	0.2%	0.5%
Home Improvement	1.0%	1.8%	0.3%	93.9%	1.8%	0.0%	0.2%	0.2%	0.4%	0.3%	0.1%
Investment	0.7%	5.9%	1.0%	0.9%	89.2%	0.3%	0.2%	0.8%	0.5%	0.4%	0.1%
Medical	1.3%	1.8%	0.8%	0.3%	2.3%	91.0%	0.0%	0.8%	0.8%	0.5%	0.5%
Moving	0.8%	1.6%	1.6%	4.0%	2.1%	0.0%	88.6%	0.0%	0.5%	0.3%	0.5%
Necessity	0.0%	5.2%	0.1%	0.4%	4.4%	0.1%	0.0%	88.7%	0.5%	0.4%	0.1%
Refinance / Payoff	0.0%	9.8%	1.1%	0.3%	1.5%	0.1%	0.0%	0.3%	86.5%	0.3%	0.1%
Vehicle	0.3%	1.4%	0.8%	0.6%	1.2%	0.1%	0.1%	0.2%	0.2%	95.1%	0.1%
Wedding	0.7%	0.9%	0.9%	0.5%	2.2%	0.0%	0.4%	0.1%	0.4%	0.3%	93.6%

Figure 8: Transition matrix of topic assignments under adversarial perturbations aimed at increasing predicted default probabilities. Each row represents the original topic; each column the reassigned topic. Diagonal values indicate the proportion of samples that retained their original topic.