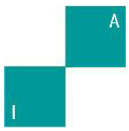
Inteligencia Artificial, 28(75) (2025), 220-247
doi: 10.4114/intartif.vol28iss75pp220-247



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in P2P Lending

Mario Sanz-Guerrero[1,2,*], Javier Arroyo[3,4]

[1]Facultad de Informática, Universidad Complutense de Madrid, Spain
[2]Johannes Gutenberg University Mainz, Germany
[3]Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, Spain
[4]Departamento de Ciencias de la Computación, Universidad de Alcalá de Henares, Spain

[*]msanzgue@uni-mainz.de

Abstract Peer-to-peer (P2P) lending connects borrowers and lenders through online platforms but suffers from significant information asymmetry, as lenders often lack sufficient data to assess borrowers' creditworthiness. This paper addresses this challenge by leveraging BERT, a Large Language Model (LLM) known for its ability to capture contextual nuances in text, to generate a risk score based on borrowers' loan descriptions using a dataset from the Lending Club platform. We fine-tune BERT to distinguish between defaulted and non-defaulted loans using the loan descriptions provided by the borrowers. The resulting BERT-generated risk score is then integrated as an additional feature into an XGBoost classifier used at the loan granting stage, where decision-makers have limited information available to guide their decisions. This integration enhances predictive performance, with improvements in balanced accuracy and AUC, highlighting the value of textual features in complementing traditional inputs. Moreover, we find that the incorporation of the BERT score alters how classification models utilize traditional input variables, with these changes varying by loan purpose. These findings suggest that BERT discerns meaningful patterns in loan descriptions, encompassing borrower-specific features, specific purposes, and linguistic characteristics. However, the inherent opacity of LLMs and their potential biases underscore the need for transparent frameworks to ensure regulatory compliance and foster trust. Overall, this study demonstrates how LLM-derived insights interact with traditional features in credit risk modeling, opening new avenues to enhance the explainability and fairness of these models.

Keywords: Credit Risk, Peer-to-Peer Lending, Natural Language Processing, BERT, Transfer Learning, Explainable AI

1 Introduction

Peer-to-peer (P2P) lending is a growing phenomenon that allows individuals to engage in direct lending and borrowing transactions, bypassing traditional financial institutions. The process is facilitated through online platforms, where prospective borrowers submit loan applications and potential lenders make informed decisions about where to invest their funds.

An inherent challenge in P2P lending is the presence of information asymmetry, wherein borrowers possess more and often superior information compared to lenders. To address this issue, platforms employ strategies to complement the conventional data provided in loan applications [12]. For instance, borrowers

ISSN: 1137-3601 (print), 1988-3064 (on-line)

are frequently encouraged to provide a voluntary textual description describing the purpose of the loan and their particular situation. Despite the absence of formal verification, such voluntary disclosures have been observed to stimulate increased bidding activity among lenders. However, lenders may lack the expertise to assess the creditworthiness of borrowers effectively and may be influenced by different factors [33].

Traditional credit scoring models often overlook the valuable information contained in loan applicants' narratives [35]. Several methods have attempted to incorporate this data, including extracting linguistic metrics [16, 15], using topic modeling to identify underlying themes [55, 62], or combining both approaches [46]. These methods offer clear advantages, such as computational efficiency—particularly in the case of linguistic metrics—and greater interpretability. However, they also come with drawbacks, including limited ability to capture meaning—especially for linguistic metrics—, dependence on extensive preprocessing in the case of topic modeling, and an overall inability to fully grasp nuance and context.

In contrast, transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) [13], offer a powerful alternative. While these models are less interpretable, they excel at understanding text by capturing semantic and contextual relationships at the word, sentence, and document levels. They also adapt flexibly to variations in style and structure. Moreover, leveraging pretrained models is straightforward—by transferring general language knowledge from large corpora, these models can be fine-tuned for specific tasks or domains, achieving significant performance improvements with relatively little labeled data.

In sum, BERT's bidirectional training and context-sensitive representations make it well-suited for tasks requiring deep semantic understanding. It has been successfully fine-tuned for various classification tasks [48], including applications in biomedicine [25] and specialized areas such as spam detection [52]. Recently, Xia et al. [57] demonstrated the effectiveness of a fine-tuned BERT model in discriminating P2P loans within the Chinese market.

In this study, we extend this line of research by applying BERT at the loan granting stage—a critical point where customer narratives hold greater importance due to the limited information available for decision-making. As our baseline method, we employ XGBoost, an efficient gradient-boosting algorithm that builds ensembles of decision trees to enhance performance and that has demonstrated its ability in loan granting scenarios [2]. We show that incorporating BERT-generated risk scores into a loan granting model significantly enhances predictive performance, and we delve deep into how BERT processes textual data and influences model behavior. Specifically, our analysis reveals that BERT captures a wide range of information from loan descriptions, including borrower-specific attributes, loan purposes, and linguistic features embedded in the text. Furthermore, we demonstrate that the inclusion of BERT-generated scores reshapes how credit models leverage other input variables, with the impact varying substantially across different loan purposes. While our findings highlight BERT's potential to improve credit risk assessment in P2P lending, they also emphasize the importance of transparency in understanding what these models learn from text. Such transparency is crucial for building trust among stakeholders and ensuring acceptance by entities.

This paper is organized as follows: Section 2 provides a comprehensive review of related work in credit risk assessment and natural language processing. Section 3 presents an overview of LLMs and the BERT model. Section 4 describes the dataset used, detailing the data preprocessing steps and conducting an in-depth exploratory data analysis. Section 5 outlines the methodology, model architecture, and training procedures employed in integrating BERT into the credit risk assessment framework. Section 6 analyzes the risk score generated by the BERT description processing. Section 7 discusses the results of our experiments, highlighting the improvements achieved by incorporating BERT-based textual analysis. Finally, Section 8 concludes the paper by summarizing key findings, discussing implications, and suggesting avenues for future research in the intersection of NLP and credit risk assessment.

2 Related Work

2.1 Data Sources in Credit Risk Modeling: The Use of Loan Descriptions

In their comprehensive analysis of risk-return modeling within the P2P lending market [4], the authors identify a discernible trend towards including new sources and types of information to improve risk and

profit management models in the P2P market. The sources are very diverse and include transactional data [59], the topology of the lending-borrowing network [27], data from social networks [58], or, more recently, facial features [38]. Among them, the authors identify as a predominant trend the inclusion of textual data taken from statements describing the purpose of the loan.

In a pioneer work exploring the impact of textual factors on peer-to-peer lending [19], the authors analyze P2P loans including manually annotated narrative aspects, such as trustworthiness, economic hardship, hard work, success, morality, and religiosity. These aspects were combined with demographic variables and loan characteristics. Their results highlight that narratives regarding trustworthiness strongly influence decision-makers, particularly credit lenders, in their loan approval process. Additionally, some of these narratives play a substantial role in subsequent loan performance.

However, most subsequent studies typically use text mining or artificial intelligence methods to extract linguistic features or loan description topics. Regarding the use of linguistic features, the authors in [16] use machine learning and text mining techniques to quantify and extract linguistic features (e.g., readability, positivity, objectivity, and deception cues), and then build both explanatory econometric models and predictive models using such features. They find that they can indeed reflect borrowers' creditworthiness and predict loan default. They also use a panel of investors and confirm that investors indeed value texts written by borrowers, but that they can also be deceived by some of the deception cues well established in the literature. Similarly, in [15], the authors include linguistic factors and the presence of social and emotional keywords and evaluate their impact on two European platforms. They found that text-derived variables influence the probability of funding, but not the probability of default. In [54], linguistic statistical features and abstract text features (including deception, subjectivity, sentiment, readability, personality, and mindset) are used to characterize text descriptions. They compare the performance of different classifiers based on the textual features and conclude that their performance is close to that of the classifiers using traditional financial features, but that adding textual features can improve the performance of the whole credit risk evaluation system.

2.2 Topic Modeling Approaches

As for topic modeling, the Latent Dirichlet Allocation model (LDA) has been widely used. In [22], the authors use LDA to extract six topics from the loan descriptions whose meanings are obvious: assets, income and expenses, work, family, business, and agriculture. They also consider the number of characters in the descriptive text. They conclude that soft (qualitative) information can improve the performance of loan default prediction compared to existing methods based only on hard (quantitative) information and that soft features have a significant ability to discriminate loan defaults. Similarly, in [60], an LDA topic model is used to classify the loan titles into six purposes. Their findings reveal that the stated purpose significantly influences a borrower's chances of securing financing. Notably, ambiguous titles—where borrowers fail to clearly articulate the loan's purpose—substantially diminish the likelihood of loan approval. In [55], Xia et al. used a keyword clustering algorithm for automatic topic extraction. Their method combines keyword extraction based on term frequency-inverse document frequency (TF-IDF) with word embeddings generated by the Word2Vec neural network model [34]. Analysis of three real-world datasets demonstrated that incorporating these topic variables significantly enhanced predictive accuracy compared to relying solely on traditional information.

Siering [46] recently examined the impact of both topical and linguistic features on loan default prediction. To extract topics, the author employed a financial text analysis method [29] to construct a domain-specific dictionary. The identified topics captured elements such as the loan purpose, the borrower's requests for assistance, expressions of reliability, and appreciation. These topics were represented as binary indicator variables. Additionally, text mining techniques were used to generate features measuring attributes like polarity, *active orientation*, readability, average sentence length, and word count. These features were then incorporated into a logistic regression model, revealing that both linguistic and content-based factors contribute to predicting loan default probability, with content-based factors showing greater significance. The analysis further indicated that certain variables, such as expressions of reliability, positively correlate with loan repayment likelihood.

2.3 Advancements with Large Language Models

In recent years, several studies have begun to explore state-of-the-art natural language processing techniques, including deep learning models, for loan default prediction. In their work [62], Zhang et al. investigated transformer encoders for extracting textual features from loan descriptions. These features, combined with traditional hard features from loan applications, were input into a neural network to predict default probabilities. Their findings demonstrated the effectiveness of transformers, with models incorporating textual loan descriptions outperforming those that did not.

More recently, Xia et al. [57] explored the use of BERT-based LLMs to extract information from loan narratives, integrating these insights into both logistic regression and machine learning models like random forest and deep forest. While this work achieved improved predictive performance, it lacked transparency regarding the specific information captured by the LLM and how it influenced the classification model, leaving the results largely opaque and black-box.

2.4 Research Gap

Despite the progress made in leveraging textual data for credit risk modeling, several gaps remain. First, while advanced NLP techniques like BERT have been applied to loan descriptions, the interpretability of the extracted features and their influence on downstream models remains underexplored. For instance, a previous work [57] achieved improved performance but provided limited insights into how the BERT-derived features contributed to the model's predictions. Second, there is a need for a more systematic integration of textual features with traditional financial variables in a way that enhances both predictive accuracy and interpretability.

Our study addresses these gaps by applying BERT to create a risk score from loan descriptions in the loan-granting process, a context where customer narratives play a crucial role. Unlike previous work, we analyze the score, including its relationship with other variables, its impact on the classification model, and on the results obtained across the predefined loan purposes segmented, aiming to provide more interpretability and insights into the decision-making process.

3 Basics of LLM architectures and BERT

Large Language Models (LLMs) are built upon the Transformer architecture [53], leveraging attention mechanisms to enhance language comprehension. LLMs can be broadly categorized into three primary families, each distinguished by its architecture:

- Encoder-only, widely employed for language comprehension tasks such as text classification, named entity recognition, and extractive question answering. The most famous example is BERT [13], which will be explained in more detail below.
- Decoder-only, designed for generative tasks, exemplified by the well-known GPT models [39, 8]. It is employed in various tasks, including question answering [36], text summarization [6], and programming code generation [10].
- Encoder-decoder models, suited for tasks demanding both language understanding and generation, such as language translation or text summarization. The most influential models are BART [26] and T5 [40].

The selection of the appropriate architecture hinges on the specific requirements of the intended task. Whether it be the nuanced comprehension of language, creative text generation, or the synthesis of both, the versatility of LLMs offers a tailored solution for diverse applications.

We will focus on BERT (Bidirectional Encoder Representations from Transformer), which is a Transformer-based language model introduced by Google researchers in 2018 [13]. BERT's architecture consists of a stack of encoders from the Transformer model. The bidirectional nature of BERT is key, as it considers both the left and right context of each word, enhancing its ability to understand context-dependent meanings and to be effective in language understanding tasks. Numerous studies have consistently shown that BERT is the most effective linguistic model for various of these tasks [23, 47]. Notably, BERT has

340 million parameters, while the widely recognized GPT-3 model has 175 billion, making BERT 514 times smaller than GPT-3 [8]. Given this significant size difference, BERT can be operated on standard home equipment for model inference, which greatly simplifies its use in practical scenarios. In contrast, GPT, built with a Transformer decoder stack, not only demands much more powerful equipment but is suited for language generation tasks.

BERT stands as a milestone whose success has spurred the development of a diverse family of models that build upon its architecture. Some versions aim to achieve similar performance while having a smaller number of parameters, such as DistilBERT (a distilled version of BERT) [44], or ALBERT (A Little BERT) [24]. Others are adaptations to other languages such as CamemBERT [32] to French or BETO to Spanish [9]. Other proposals aimed to improve upon BERT by modifying some design decisions when pretraining BERT and also training the model longer, as in the case of RoBERTa (Robustly optimized BERT approach) [28], which resulted in improved contextualized representations and enhanced language understanding.

To further elucidate the role of BERT in specialized applications, it is crucial to understand its capacity for transfer learning and fine-tuning. Transfer learning involves using a pre-trained model like BERT, which has initially learned general language patterns from a large corpus to a specific task or dataset. This technique allows us to take advantage of the rich linguistic representations without needing extensive computation from scratch. Fine-tuning involves adjusting the pre-trained model's parameters to capture the nuances of the target task or application field by further training with new instances from the new context. For example, BioBERT is a BERT model fine-tuned for biomedical text mining tasks like named entity recognition and question answering [25]. Other adaptations have targeted text classification and sentiment analysis in specific datasets [48, 17]. Finally, as already mentioned, [57] shows how fine-tuned Chinese BERT models can enhance the classification performance of default loans in the P2P market.

4 Dataset

We use a public data set of the P2P lending company Lending Club[1], which is widely used in credit risk publications and the most widely used when dealing with the P2P market [1, 4]. However, instead of using the original dataset, which includes 2,260,699 loans granted by the company between 2007 and 2018, we use a version modified for proposing granting models [5], used in [2, 3]. Since granting models determine which loans will be fully repaid, its estimation needs loans whose final status is known (i.e., that were either fully repaid or defaulted). Thus, the dataset excludes loans in transitory states (in a grace period, late, etc.) and loans with no information on income and indebtedness, which is essential to compute the input variables, resulting in 1,347,681 instances.

Additionally, the original dataset contains variables detailing the loan's lifecycle and other post-application aspects (e.g., the interest rate). In contrast, our version only includes variables available at the time of application, which are those utilized by granting models.

Loan descriptions were inconsistently available, appearing only for certain loans between April 2008 and March 2014. To accurately assess the impact of textual descriptions on default prediction, our analysis focuses solely on the 119,101 loans that include the *desc* variable. Kolmogorov-Smirnov and chi-square tests were applied to quantitative and categorical variables to assess potential bias from filtering. The lack of significant differences indicates that the filtered dataset is representative of the original dataset.

In the dataset, the target variable suffers the usual class imbalance problem (only 15.27% of default), which will be considered in the design of the experiments. Table 1 shows the input variables of our granting model, which are explained below.

As for the quantitative variables, the Fair Isaac Corporation credit bureau (FICO) information in the original dataset is given by a minimum and maximum range of limits to which the borrower's FICO belongs at loan origination. However, we average these two values to have a single indicator of the creditworthiness of potential borrowers resulting in our *fico_n* variable. For the case of the debt variable, *dti_n* is estimated from the original dataset variables as the ratio calculated from the total debts of the co-borrowers over the total debt obligation divided by the combined monthly income of the co-borrowers.

[1] `https://www.kaggle.com/wordsforthewise/lending-club`

Table 1: Variable description.

Variable	Description
Quantitative variables	
revenue	Borrower's self-declared annual income during registration.
dti_n	Indebtedness ratio for obligations excluding mortgage. Monthly information.
loan_amnt	Amount of credit requested by the borrower.
fico_n	Credit bureau score. Defined between 300 and 850, reported by Fair Isaac Corporation as a summary risk measure based on historical credit information reported at the time of application.
Categorical variables	
emp_length	Employment length of the borrower categorized into 12 categories, including the no information category.
purpose	Credit purpose category for the loan request.
home_ownership	Homeownership status provided by the borrower.
addr_state	Borrower's residence state from the USA.
Textual variable	
desc	Description of the credit request provided by the borrower.

Regarding the categorical variables, we merged the categories 'other', 'none', and 'any' into a unified category labeled 'other' for the *home_ownership* variable. This decision was made due to a lack of clear differentiation among these options, coupled with their similar default percentages and their relatively low percentages of occurrences. The *emp_length* variable was treated as categorical rather than numerical since it includes categories for 'no information' and for 'more than ten years'.

For the textual variable, we carried out an exhaustive work of text cleaning. First, we removed all those descriptions that contained the default description provided by Lending Club on its web form (*"Tell your story. What is your loan for?"*). Moreover, we removed the prefix *"Borrower added on DD/MM/YYYY >"* from the descriptions, as we did not want any temporal background on them. Finally, as these descriptions came from a web form, we replaced all HTML entities with their corresponding characters (e.g. '&' was substituted by '&', '<' was substituted by '<', etc.).

Table 2 presents the quantitative variables and the results of the Kolmogorov-Smirnov test, which was used to compare the empirical cumulative distribution functions of Default and Non-default loans. According to these results, defaulted loans are characterized by lower revenue, higher debt-to-income ratio (*dti_n*), higher requested amount (*loan_amnt*), and lower FICO scores (*fico_n*), being the differences significant at the 0.01 level.

Similarly, Table 3 displays the distribution of categories within each categorical variable and the corresponding default rates. The *addr_state* variable is excluded due to its 50 categories, one for each U.S. state. The table also indicates whether there is a significant dependence between the target variable and the categorical variables at the 0.01 significance level. The results show significant dependence for all variables, including the *addr_state* variable not reported in the table (test value of 211.12).

In the *home_ownership* variable, the 'OTHER' category shows the highest risk (20.98%) but a small frequency (0.12%), while the 'MORTGAGE' category is the most frequent (51.05%) and the least risky (14.14%) one. In the *emp_length* variable, the category that denotes no information ('NI') has the highest risk (19.78%), but also the lowest frequency (4.13%). In general, employment length can be categorized into two groups with comparable default rates: those with employment lengths of five years or less and those with more than five years. Interestingly, the risk is slightly higher in the group with more than five years of employment. The categories are not perfectly ordered, which supports the use of one-hot encoding to treat this variable as categorical. Finally, the most frequent *purpose* is 'debt consolidation', constituting 57% of the loans, which has a default rate of 16.14%. Notably, the riskiest purpose is 'small business', with a 26.41% default rate. Conversely, 'car' loans demonstrate the lowest risk, with a mere 9.61% default rate. This striking divergence in default rates across diverse purposes underscores a significant variability in risk within the various loan purposes.

Regarding the textual description of the loan (*desc* variable), Table 4 shows some metrics to charac-

Table 2: Exploratory data analysis. Quantitative variables.

Variable	Statistic	Non-Default	Default	Total
revenue	Mean	$ 73,570.69	$ 66,218.66	$ 72,447.84
	Median	$ 64,000.00	$ 58,000.00	$ 62,000.00
	SD	$ 54,944.60	$ 40,731.98	$ 53,086.79
	KS D-Test			0.09*
dti_n	Mean	16.15	17.67	16.38
	Median	15.88	17.70	16.16
	SD	7.50	7.53	7.53
	KS D-Test			0.09*
loan_amnt	Mean	$ 13,799.25	$ 15,111.44	$ 13,999.66
	Median	$ 12,000.00	$ 14,000.00	$ 12,000.00
	SD	$ 7,931.55	$ 8,363.19	$ 8,012.86
	KS D-Test			0.08*
fico_n	Mean	705.22	694.20	703.54
	Median	697.00	687.00	697.00
	SD	33.32	27.48	32.74
	KS D-Test			0.15*

* p-value less than 0.01.

terize it. There is a one-word difference in the average word count between the descriptions of defaulted and non-defaulted obligations. The readability was calculated using the Flesch Reading Ease Score[2], which indicates the approximate educational level required for comfortable comprehension of a given text (higher scores denote greater ease of reading). The texts in both categories have scores around 66, signifying that they can be readily comprehended by students aged 13 to 15. Additionally, we analyzed the average polarity and average subjectivity[3]. The polarity, ranging from -1 to 1 to denote negative or positive sentiment, was observed to be approximately 0.1 in both cases, suggesting a subtle positive sentiment. On the other hand, subjectivity, measuring the presence of judgments and opinions on a scale from 0 to 1, exhibited values close to 0.31 in both categories. This indicates that while the texts in both cases maintain a generally objective tone, there is a discernible inclusion of some judgments or opinions.

Although the distinctions in these metrics between the default and non-default categories are subtle, their significance is confirmed by the Kolmogorov-Smirnov test. Consequently, it is pertinent to incorporate linguistic aspects into credit risk modeling. Our approach for extracting information from the descriptions relies on leveraging LLMs capable of encompassing not just linguistic nuances but also capturing content details. We elaborate on our methodology below.

5 Methodology

This study employs transfer learning to enable an LLM to generate a score that reflects the likelihood of loan default based on textual descriptions. We explore how a fine-tuned LLM captures various aspects of loan descriptions that are indicative of default risk. Furthermore, we demonstrate that incorporating the LLM-generated score enhances the predictive accuracy of a loan-granting model and fundamentally alters how the model operates compared to when the score is absent.

Our baseline model is a machine learning classifier that utilizes all available variables from the loan application process, including both quantitative and categorical data. Specifically, we use XGBoost,

[2]Calculated with Textstat (Python library). Source: `https://github.com/textstat/textstat`.
[3]Calculated with TextBlob (Python library). Source: `https://github.com/sloria/textblob`.

Table 3: Exploratory data analysis. Categorical variables.

Variable	Category	Count	Rel. Freq.	Default Rate	Chi Test
home_ownership	MORTGAGE	60,796	51.05%	14.14%	131.08*
	OTHER	143	0.12%	20.98%	
	OWN	9,582	8.05%	15.69%	
	RENT	48,580	40.79%	16.60%	
emp_lenght	< 1 year	9,548	8.02%	14.83%	104.96*
	1 year	7,803	6.55%	14.39%	
	2 years	10,960	9.20%	14.68%	
	3 years	9,370	7.87%	14.18%	
	4 years	7,561	6.35%	14.56%	
	5 years	9,019	7.57%	14.76%	
	6 years	7,271	6.10%	16.04%	
	7 years	6,638	5.57%	15.59%	
	8 years	5,374	4.51%	15.39%	
	9 years	4,356	3.66%	15.79%	
	10+ years	36,287	30.47%	15.41%	
	NI	4,914	4.13%	19.78%	
purpose	car	1,884	1.58%	9.61%	568.47*
	credit card	25,051	21.03%	12.81%	
	debt consolidation	68,372	57.41%	16.14%	
	educational	265	0.22%	16.98%	
	home improvement	7,170	6.02%	12.93%	
	house	805	0.68%	15.78%	
	major purchase	3,062	2.57%	10.65%	
	medical	970	0.81%	17.32%	
	moving	768	0.64%	14.84%	
	other	6,361	5.34%	17.69%	
	renewable energy	127	0.11%	19.69%	
	small business	2,518	2.11%	26.41%	
	vacation	561	0.47%	16.22%	
	wedding	1,187	1.00%	12.47%	

* p-value less than 0.01.

Table 4: Exploratory data analysis. Textual variable (*desc*).

Variable	Statistic	Non-Default	Default	Total
Word count	Mean	36.72	35.49	36.54
	Median	24.0	22.0	24.0
	SD	46.62	48.78	46.96
	KS D-Test			0.03*
Readability	Mean	66.70	66.31	66.64
	Median	73.88	74.19	74.02
	SD	32.87	35.55	33.29
	KS D-Test			0.02*
Polarity	Mean	0.0964	0.0909	0.0956
	Median	0.0367	0.0	0.0320
	SD	0.1685	0.1699	0.1687
	KS D-Test			0.04*
Subjectivity	Mean	0.3193	0.3029	0.3168
	Median	0.3635	0.3333	0.3589
	SD	0.2542	0.2595	0.2551
	KS D-Test			0.04*

* p-value less than 0.01.

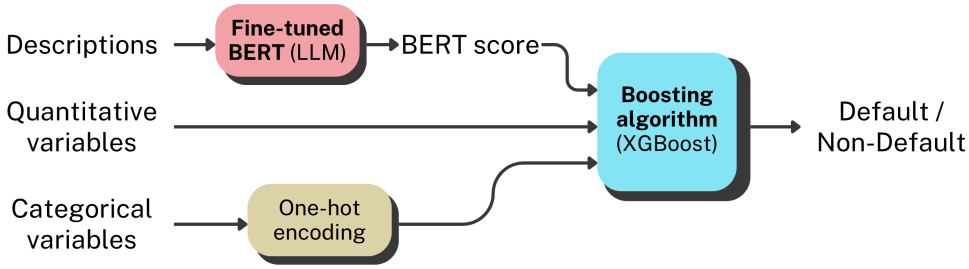


Figure 1: Diagram of the experiment architecture.

which has been shown to deliver superior performance in similar loan-granting contexts [2]. To augment this baseline, we introduce the LLM-generated default risk score as an additional input feature. We analyze the information captured by this score and assess its impact on both prediction performance and the behavior of the resulting model.

The experimental setup is shown in Figure 1. Loan descriptions are processed by a fine-tuned BERT model, which outputs a *BERT_score* representing the probability of default. This score is then integrated with other input variables in the XGBoost classifier to produce the final prediction.

The key components of the methodology are elaborated upon in this section.

5.1 Tuning the classifier

The classification algorithm used was XGBoost [11], which is trained using a stratified k-fold cross-validation, dividing the dataset into k subsets (here, $k = 5$) and preserving the original class distribution in each fold to provide a more reliable evaluation in an imbalanced dataset as ours. The dataset is shuffled to eliminate biases derived from its original ordering and to ensure representative subsets, avoiding artificial patterns in the training. Furthermore, k=5 is used to balance the reduction of variance in the estimates

Table 5: Hyperparameters of XGBoost considered in the genetic optimization and their respective ranges.

Parameter	Min. value	Max. value
scale_pos_weight	0.1	10
eta (learning rate)	0.001	0.5
subsample	0.7	1
n_estimators	2	500
colsample_bytree	0.3	1
max_depth	2	12
lambda	0.5	10
alpha	0.5	10
gamma	0	10
min_child_weight	0	10

with the need for sufficiently large partitions, a crucial factor in datasets with few instances and high heterogeneity as ours.

Furthermore, the instances are shuffled to avoid potential ordering biases in the dataset. To fine-tune the hyperparameters we used a genetic algorithm [20] to evolve candidate hyperparameter combinations and choose the one that maximizes the fitness measure, which was the average balanced accuracy (BACC) in the 5 validation sets of the cross-validation. We use BACC as it accounts for the imbalanced nature of the dataset. In preliminary experiments, we also considered the area under the receiver operating characteristic (AUROC), which measures the model's ability to discriminate between positive and negative examples regardless of the classification threshold chosen. However, we observed that the resulting XGBoost classifiers produced poor BACC values (similar to those from a naïve classifier that predicts the majority class) when using the standard 0.5 threshold to make the prediction. We also observed that XGBoost classifiers with extremely similar AUROC values produced very different results in terms of BACC. Thus, we decided to use the BACC measure as it resulted in classifiers with more stable behavior.

In genetic optimization, each individual is characterized by its genes, that is, the considered hyperparameters of the XGBoost. We have included several kinds of parameters, including:

- Parameters that adjust the sample weights, such as *scale_pos_weight*, which balance the weights of the classes and are useful in unbalanced datasets as ours.

- Parameters that set up the behavior of the boosting algorithm, such as the learning rate (*eta*), the percentage of subsamples in each iteration (*subsample*), the number of learners (*n_estimators*), and the percentage of dataset features that use each learner (*colsample_bytree*).

- Parameters that control the learning process of each tree, including its maximum depth (*max_depth*), regularization parameters (*lambda* and *alpha*), the loss reduction required to make a further partition on a leaf node (*gamma*), and the minimum number of weighted instances needed in a child node (*min_child_weight*).

Table 5 shows the hyperparameters together with their respective ranges, which were determined based on a combination of domain knowledge, preliminary experiments, and established practices in the literature [2, 61]. For instance, the learning rate (*eta*) was set between 0.001 and 0.5 to balance the trade-off between convergence speed and model performance. The *max_depth* parameter, which controls the maximum depth of a tree, was set between 2 and 12 to prevent overfitting while allowing sufficient model complexity. Similarly, the ranges for other hyperparameters, such as regularization terms and sampling rates, were selected to ensure a wide exploration of their effects, which are crucial for optimizing model performance in imbalanced datasets.

The evolutionary strategy chosen for the optimization is the "*Mu* plus *lambda*" ($\mu + \lambda$) approach, where μ represents the number of individuals to select for the next generation, and λ indicates the number of children to produce at each generation. Unlike traditional approaches where children often replace parents, the $\mu + \lambda$ strategy involves adding both children and parents to produce the next generation. This strategy was selected to maintain diversity in the population and prevent premature convergence

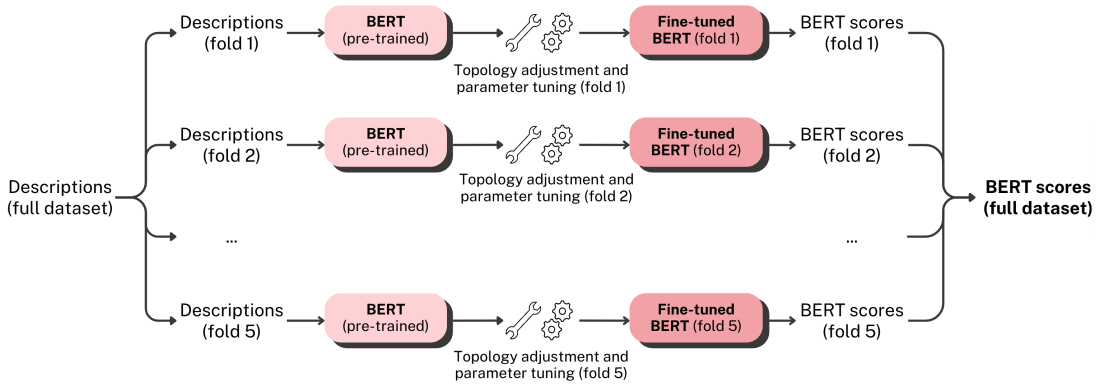


Figure 2: Fine-tuning process of BERT for generating a default score from loan descriptions.

to suboptimal solutions. In the context of this research, we set μ to 150 and λ to 150. This configuration, chosen based on empirical testing, provides a balance between exploration of the search space and computational feasibility. Increasing μ and λ beyond these values resulted in marginal improvements at significantly higher computational costs.

Initially, pairs of parents are chosen through tournament selection with a tournament size of 2. Subsequently, the children are generated employing a two-point crossover technique on the parents' chromosomes with an 80% probability and applying a random resetting mutation with a 20% probability. The random resetting mutation implies that each gene of every child has a 20% chance of acquiring a new random value within its defined range. These probabilities were chosen based on preliminary experiments that demonstrated a good balance between exploration and exploitation. To create an offspring of 150 children, this selection, crossover, and mutation process is repeated 75 times. Finally, we combine the 150 children and the 150 parents, resulting in generations of 300 individuals, and select the top 150 (μ) according to their fitness value to pass to the next generation.

The evolutionary process consists of 20 iterations, thereby generating a total of 3,000 individuals—each representing a distinct hyperparameter configuration. The choice of 20 iterations was determined based on convergence analysis, where we observed that fitness improvements plateaued after approximately 15–20 generations. From this pool of configurations, the one exhibiting the highest fitness is ultimately chosen as the optimal outcome.

5.2 Generating a default score with BERT

In this section, we delineate the methodology employed to generate a default score based on the textual description of the loan. We initiate the process by applying transfer learning utilizing an LLM, specifically BERT in our case. The fine-tuning of BERT, illustrated in Figure 2, results in a model that produces an outcome within the range of 0 to 1, offering a nuanced indicator rather than a binary classification. This subtle indicator is subsequently integrated into a classifier along with other input variables to predict the likelihood of loan default. The subsequent steps in this process are detailed next.

5.2.1 Transfer learning to produce the default score

The BERT model we utilize[4] is configured with L=12 hidden layers (i.e., Transformer encoder blocks), each with a size of H=768, and it employs A=12 attention heads. These attention heads enable the model's self-attention mechanism to process inputs in 12 distinct patterns simultaneously. The output from BERT are embeddings of size 768. For incorporating this model, TensorFlow HUB was selected for its efficient integration with additional neural network layers.

As outlined in Section 3, during the transfer learning phase, we aim to exploit BERT's advanced language understanding while minimizing the need to learn from scratch. To achieve this, we freeze the weights of all but the last hidden layer of BERT. This approach preserves the model's pre-trained

[4]Source: `https://tfhub.dev/tensorflow/bert_en_uncased_L-12_H-768_A-12/4`

capabilities and prevents overfitting on our dataset, while also mitigating catastrophic forgetting—a phenomenon where neural networks lose previously acquired knowledge when retrained on new tasks [7].

Subsequently, in the fine-tuning stage, only the last BERT layer and the newly added layers are adjusted to better serve our specific task of generating a default score. These layers are expected to enhance the model's adaptability to our particular requirements, allowing slight parameter adjustments for improved task-specific performance, while maintaining the general language understanding gained from BERT's initial pre-training. This strategy effectively balances specialized learning with the retention of valuable pre-trained knowledge.

To ensure that configurations are chosen based on empirical evidence, we incorporate architectural features into the training process. This allows the performance metric—in our case, balanced accuracy—to objectively guide model selection and minimize potential bias from manually imposed configurations. The parameters defining the extra layers' architecture play a crucial role in balancing the model's learning capacity and generalization ability—where, for instance, the second layer enhances learning, and the dropout layer promotes generalization. Thus, we explore various parameter configurations to determine the one that produces the best results. Specifically, we explore the 126 configurations that result from combining the following options:

- Using a first extra dense layer of 128, 256, or 512 neurons.
- Using or not a second extra dense layer of 128 neurons.
- Adding a dropout layer before or after all the extra layers.
  - Considering a dropout percentage of 0%, 0.10%, 0.20% or 0.30% for all the dropout layers.
- Using a learning rate of 0.001, 0.0001, or 0.00001.

All the internal hidden layers are dense layers using the $ReLU$ activation function[5]. Additionally, to obtain the probability of belonging to the default class, the last layer of the neural network was configured with a single neuron and a sigmoid activation function[6].

The neural network is trained to predict the loan outcome (default or non-default) based on the textual description as input. As a loss function, we use the weighted binary cross-entropy, which quantifies the difference between the predicted probabilities and the true binary labels by assigning different weights to each class. This approach ensures that the model does not bias predictions towards the majority class in imbalanced datasets, as it penalizes errors on the minority class more heavily. Considering class weights is crucial given the imbalanced nature of our dataset, where defaulted loans constitute only 15.27% of the total instances. The training was set up with a batch size of 64 and trained for 25 epochs, with an early stopper of 3 epochs.

The fine-tuning of the BERT model was performed on a system equipped with an Intel Core i9-12900KS processor, an NVIDIA GeForce RTX 4090 graphics card (24GB), and 128 GB of RAM. In contrast, the inference process with the fine-tuned model, which involves generating the BERT score for a given loan description, can be conducted on a standard PC. This capability enhances the practicality of deploying our approach in real-world applications.

5.2.2 Avoiding data leakage in cross-validation

As previously mentioned, our BERT model generates a default probability, which is then integrated into the classifier as an additional quantitative variable. It is important to note that when computing the BERT default probability, we must replicate the exact folds used in the k-fold cross-validation of the boosting algorithms. In each iteration, boosting algorithms are trained using the *BERT_score* variable. It is essential to prevent BERT from training with validation data from a specific fold to avoid distorting the model's true performance. Allowing this would incorporate BERT predictions from its training phase

[5]The Rectified Linear Unit (ReLU) activation function outputs the maximum of zero and the input value, "activating" the neuron if the input is positive.

[6]The sigmoid activation function introduces non-linearity and maps the input values to a range between 0 and 1, facilitating binary classification tasks.

Table 6: BERT optimal configuration and score by fold.

	Neurons in 1st dense layer	Use 2nd dense layer (128 neurons)*	Dropout	Learning rate	Loss**
Fold 0	512	True	20%	0.0001	0.1774
Fold 1	512	True	0%	0.00001	0.1793
Fold 2	128	True	20%	0.0001	0.1776
Fold 3	256	True	0%	0.001	0.1775
Fold 4	512	True	0%	0.001	0.1773

* Values 'True' or 'False' indicate whether the optimal configuration has or has not that layer.
** Weighted binary cross-entropy of the test set predictions.

in the fold's validation data, which doesn't accurately represent the model's real-world default prediction ability.

To avoid this data-leakage problem, we use the data as shown in Figure 3. This diagram consists of five steps:

1. Description extraction: Textual descriptions are extracted from the original dataset.

2. Folds generation: The exact same folds as in the boosting algorithms are generated for the textual descriptions. This is done by dividing the data into a train set (green color) and a test set (purple color). Each fold gets a different test set.

3. Optimization of the neural network architecture: The train set obtained in the previous step is divided into a 70% train subset (light green color) and a 30% test subset (dark green color). The neural network is trained with the previously mentioned configurations on the training subset (light green color), and is tested by predicting the test subset (dark green color).

4. Default prediction: The optimal configuration obtained in step 3 (lower value of weighted binary cross-entropy in the test subset) is trained with the train and test subsets (light and dark green colors) and is used to predict the default probabilities of the test set obtained in step 2 (purple color). These predictions of unseen data will be the *BERT_score* values of the current fold.

5. *BERT_score* integration: Once the *BERT_score* values of all the folds are generated in step 4, they are incorporated into the original dataset as a quantitative variable.

5.2.3 Result of the transfer learning process

As explained before, we explored 126 neural network configurations in each of the 5-fold cross-validation. The resulting optimal configuration for each validation fold and its loss score are shown in Table 6. Interestingly, the optimal combination of parameters varies across the different folds, and the only parameter that remains constant is the use of the second dense layer. Despite this parameter heterogeneity, the test loss values are quite stable, except for the case of fold 1, which has a slightly higher loss.

6 Analysis of the BERT Score

6.1 Assessment of the BERT score as a credit risk score

In this section, we evaluate the effectiveness of the BERT score in utilizing loan descriptions to predict default risk. Table 7 presents loan descriptions with the highest and lowest BERT scores, which indicate loans assessed by BERT as having the highest and lowest default risks, respectively. It appears that descriptions with higher BERT scores are often less informative, containing errors such as typos (e.g., *"pay of"*, *"everthing"*, *"alway"*, *"belive"*) or grammatical issues. Interestingly, only one of these loans ultimately defaulted. Conversely, loan descriptions with lower BERT scores tend to be more detailed and

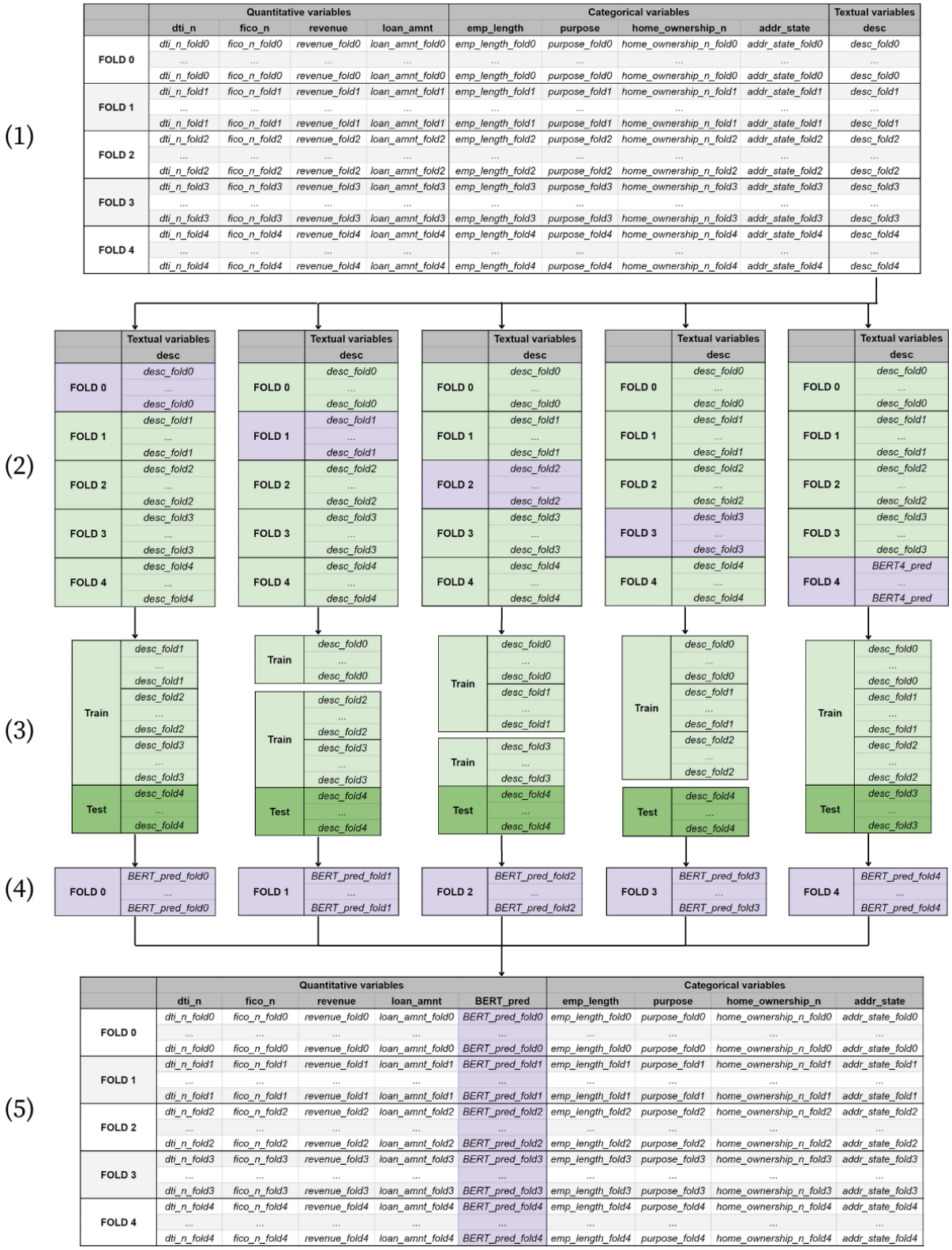


Figure 3: Data handling strategy to prevent data leakage in model training and evaluation.

Table 7: Loan descriptions with highest (top) and lowest (bottom) BERT score.

BERT score	Real value	Description
0.8562	0 (Non-Default)	*getting a divorce need new apt. with new furniture because she getting everthing.*
0.8149	1 (Default)	*need help my bills. to help pay my medications and some bills.*
0.8131	0 (Non-Default)	*i can pay of some bills for my self because i been helping other people out. i could save more for my family and their need.i have a good job that i am bless with. i am from a large family seem like every one thinking i suppose to help them when i need help my self.i alway belive that the lord will.*
0.8051	0 (Non-Default)	*consolidating our debt makes our life easier live in our means with one solid low monthly payment insted of multiple payment that add up more then what ill be paying with this loan and have a little left to leave in my savings for a rainy day to be honest and thank you for your consideration good dy.*
0.8045	0 (Non-Default)	*To consolidate debt. to pay off dept*
0.0735	0 (Non-Default)	*Debt consolidation with a lower APR.*
0.1146	0 (Non-Default)	*In need of funds to pay off some bills as well as minor improvements to house and yard. I have an extremely secure career, and maintaining my credit worthiness is important to me.*
0.1262	0 (Non-Default)	*Hard working individual with a stable job will use loan proceeds to consolidate outstanding credit cards balances. 1) Net monthly income - $4,432. 2) All expenses (allocated):. Rent - $1,124. Utilities- 84. Groceries 293. Auto (including fuel) 201 . Cell Phone 52. Cable/Internet 64. Personal care items 82. Entertainment/dining 93. Sales tax 65. 3) Previously answered. 4) No.*
0.1360	0 (Non-Default)	*This loan is to pay off credit ca.*
0.1871	1 (Default)	*I need money for moving expenses and for a buffer for the first month while I transition into working in my new location. I have successfully paid off two previous Lending Club loans in the past couple of years.*

Table 8: Classification performance of BERT binarization at 0.5.

		Default			Non-default		
Model	BACC	Precision	Recall	F1	Precision	Recall	F1
BERT	0.5444	0.1714	0.6896	0.2746	0.8771	0.3993	0.5487

insightful, often including information about the loan's purpose and the borrower's creditworthiness, and sometimes even providing numeric data related to the borrower's financial status.

Figure 4 illustrates the distribution of default and non-default loans (Y axis) across the BERT score range (X axis) in bins of 0.01. In the bar chart, the blue segment represents the percentage of non-default loans, while the orange segment denotes the percentage of default loans[7]. The figure reveals a general trend where higher BERT scores are associated with a higher proportion of defaulted loans. It is important to note that observations outside the BERT score range of $[0.3, 0.7]$ are sparse, which makes the bars in these regions less reliable. Overall, the trend suggests that higher BERT scores are indicative of a greater likelihood of default, highlighting the BERT score's usefulness as a risk assessment tool.

Table 8 shows the classification performance obtained by applying a 0.5 threshold to binarize the BERT score, which obtains a balanced accuracy of 54.4%. The BERT score is not very precise in predicting the default class (17.1%) but retrieves 69% of the instances.

These results demonstrate that the BERT model can be effectively fine-tuned to predict the final state of the loan using only the information provided by the borrower in the description field of the application form. In Section 7, we will contextualize these findings by comparing them with the results obtained from XGBoost and various sets of variables.

6.2 Relationship between the BERT score and other variables

We now explore potential relationships between the BERT score and other variables. Table 9 presents the correlation coefficients between the BERT score and the quantitative variables within the dataset.

[7]Absence of a bar indicates that no loan descriptions fall within that BERT score range.

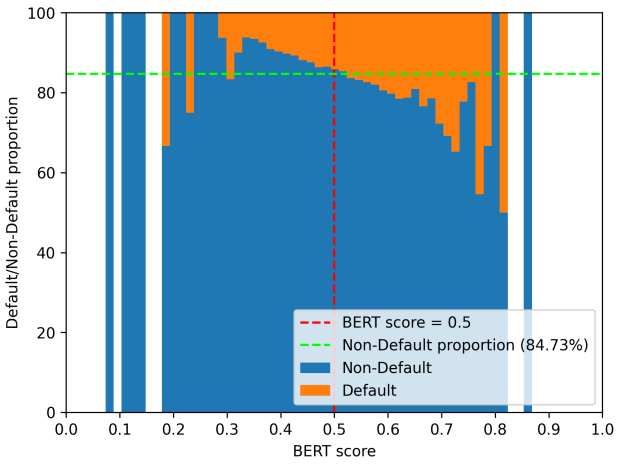


Figure 4: Default and non-default proportions by BERT score range.

Table 9: Correlation coefficients of quantitative variables with the BERT score.

Variable	Pearson	Spearman
revenue	-0.0734*	-0.1007*
dti_n	0.0663*	0.0627*
loan_amnt	-0.0008	-0.0012
fico_n	-0.1293*	-0.1293*

* p-value less than 0.01.

The findings reveal weak yet statistically significant relationships with all variables except for the loan amount. Notably, the most pronounced correlations exist with the FICO score and revenue variables. Both exhibit inverse relationships, indicating that individuals with higher FICO scores and revenues tend to have lower BERT scores, and vice versa. The association between the BERT score and the debt variable (*dti_n*) is direct, albeit slightly weaker than the other two correlations.

To evaluate the relationship with the categorical variables, Table 10 presents the results of the Kruskal-Wallis test, a non-parametric statistical test that analyzes whether there are differences in the BERT scores across categories within each categorical variable. The results indicate significant differences in BERT scores among all categorical variables, suggesting a certain level of association between the BERT score and these categorical factors. However, it remains challenging to quantify the strength of this relationship or identify the specific categories with the most robust associations.

Finally, Table 11 examines the potential relationship between the BERT score and various linguistic features automatically extracted from the text. We find statistically significant correlations between the BERT score and all linguistic features analyzed. Notably, there is a strong inverse correlation with both word count and subjectivity, indicating that shorter and more objective texts tend to have higher BERT scores. This finding, together with the correlations presented in Table 9, suggests that BERT is more closely related to linguistic features than to the numerical variables associated with loan applications.

Table 10: Kruskal-Wallis test of categorical variables with the BERT score.

Variable	H-statistic
emp_length	1581.67*
purpose	1357.94*
home_ownership	243.20*
addr_state	360.44*

* p-value less than 0.01.

Table 11: Correlation coefficients of linguistic features with the BERT score.

Variable	Pearson	Spearman
Word count	-0.1961*	-0.2650*
Polarity	-0.1036*	-0.1473*
Subjectivity	-0.1753*	-0.1866*
Readability	0.1182*	0.1518*

* p-value less than 0.01.

Table 12: Classifier performance with and without the BERT score.

Metric	Quant. + Categ. var.	Quant. + Categ. + 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

7 Results of the LLM-Enhanced Granting Model

7.1 Analysis of the classification performance

First, we evaluate the impact of incorporating the BERT score in the granting model. In our baseline experiment, we optimize XGBoost with a genetic algorithm using the quantitative and categorical variables typically used in granting models, while in the competing approach, we optimize it but include the BERT score as an input variable. Table 12 shows the results of both approaches.

A closer examination of the balanced metrics reveals a marginal enhancement in both BACC (0.6154 vs. 0.6187) and AUC (0.6575 vs. 0.6644), the latter significant according to the DeLong test [50]—a non-parametric approach for evaluating whether differences between AUCs of two models are statistically significant— at the 0.01 level. It is crucial to note that the Lending Club dataset exclusively consists of approved loans. This fact poses a substantial challenge to significantly enhance the outcomes in a loan granting model such as ours since the loans included in the dataset were initially considered favorable by the platform. Furthermore, in experiments not reported in the paper we used CatBoost [37] instead of XGBoost and obtained a similar BACC improvement.

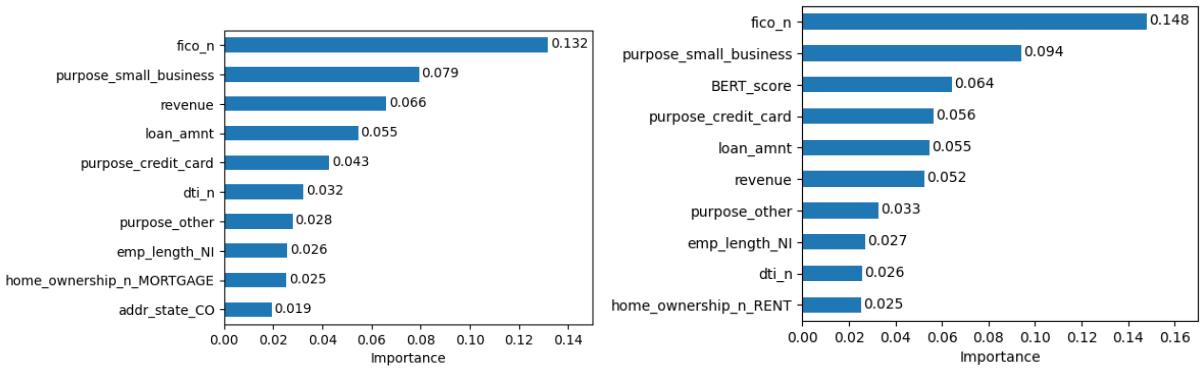
The performance metrics in Table 12 indicate that incorporating the BERT score results in improved precision but diminished recall. However, attributing this change in precision-recall behavior solely to the BERT score requires careful consideration. This caution arises from our observations within our dataset, where classifiers with different hyperparameters and similar near-optimal balanced accuracy values have demonstrated varying precision-recall behaviors, suggesting that this may also be the case here.

Table 13 shows an additional experiment in which XGBoost classifiers are trained and optimized using only one kind of input variable. While the classifier using the quantitative variables is clearly the best, the classifier that uses just the BERT score obtains slightly better results than the one using the four categorical variables (the AUC difference is significant at 0.01 according to the DeLong test). This is noteworthy given the well-known effectiveness and the meaningful nature of the qualitative variables.

Table 13 also shows the result of an XGBoost classifier using the textual features presented in Table 11, namely: polarity, subjectivity, word count, and readability score. This classifier is outperformed by the XGBoost that uses the BERT score in terms of balanced accuracy and AUC (significant at the 0.01 level according to the DeLong test). This finding underscores the superior ability of the fine-tuned LLM to leverage textual descriptions and extract relevant information for the classification task.

Table 13: Performance of the XGBoost considering only a subset of variables.

Metric	Quant.	Categ.	Text.	BERT score
BACC	0.6062	0.5486	0.5258	0.5490
AUC	0.6457	0.5656	0.5309	0.5714
F1	0.3192	0.2759	0.2534	0.2601
Precision	0.2138	0.1746	0.1665	0.1877
Recall	0.6300	0.6563	0.5302	0.5153
Accuracy	0.5896	0.4738	0.5227	0.5724



(a) Before BERT score incorporation.

(b) After BERT score incorporation.

Figure 5: Feature importance of the two XGBoost classifiers.

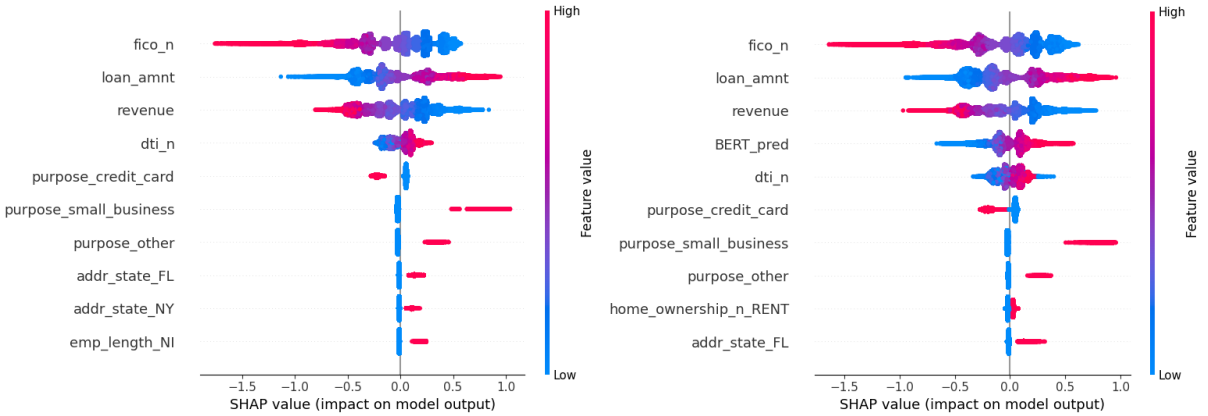
7.2 Feature importance and explainability

Now we delve into how the inclusion of the BERT score alters the classifier's use of input variables. Our goal is not only to assess the predictive contribution of the BERT score but also to understand its broader impact on the model's decision-making process. Figure 5 shows the feature importance assigned by XGBoost, both with and without the BERT score. Notably, the BERT score becomes the third most influential variable, accounting for 6% of the total importance. Moreover, its inclusion reshapes the importance and ranking of other features. For instance, the importance of the borrower's annual revenue and the debt-to-income ratio (*dti_n*), both crucial for assessing the borrower's economic status, significantly decreases. This shift suggests that the BERT score may encapsulate information that overlaps with these variables or renders them less critical, as supported by the correlation analysis in Table 9.

To analyze whether the relationships between each variable and the predicted outcome change, we use the SHAP values [30], which quantify the contribution of each feature to the model's predictions. Figure 6 compares the SHAP values for the 10 most impactful features in models both with and without the BERT score. It reveals that the distributions of SHAP values, as seen in the violin plots, do not significantly change when the BERT score is included. Interestingly, the SHAP value distribution in Figure 6a mirrors a similar analysis presented in [2], conducted on a lending model with Lending Club data[8], which aligns with expectations in the credit risk context.

Figure 7 reveals a direct and linear relationship between the BERT score and the SHAP values, which in this case relate to the default risk. Notably, this relationship is asymmetric around BERT score values of 0.5; scores below 0.4 correspond to SHAP values ranging between -0.4 and -0.6, strongly guiding the model toward predicting non-default. Conversely, this impact range in the positive case is only reached by BERT scores exceeding 0.7, signifying that only exceptionally high BERT scores serve as strong indicators of default.

[8]Our dataset is narrower, considering only loans accompanied by textual descriptions, as detailed in Section 4.



(a) Before BERT score incorporation. (b) After BERT score incorporation.

Figure 6: SHAP values of the XGBoost classifiers.

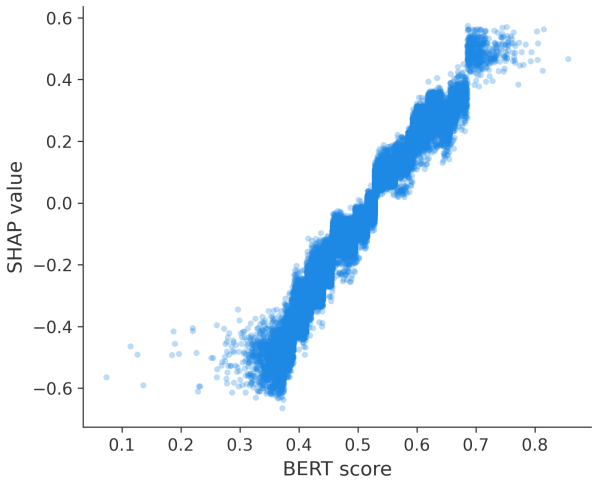
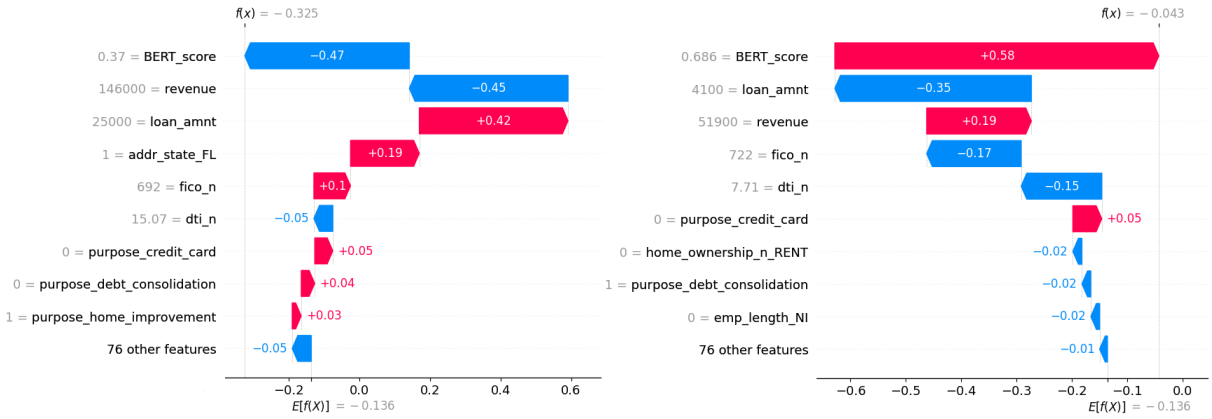


Figure 7: Dependence plot between SHAP values and BERT score.

7.3 A closer examination of the role of the BERT score in classification

As shown in our analysis of feature importance in XGBoost, the BERT score plays a crucial role in the correct classification of the loans. Using SHAP values, we can determine how each variable influences the classification decision for a given loan. Figure 8 shows the waterfall plot for two loans where the BERT score is a decisive factor. The plot should be read from bottom to top. It begins at the bottom with the expected value of the model output. Each row then shows how each feature's positive (red) or negative (blue) contribution shifts the prediction from this expected value to the final prediction for that specific loan. Features are ordered from bottom to top in ascending order of contribution (in absolute value). The x-axis represents log-odds units (the margin output before the logistic link function used by the classifier), meaning that negative values correspond to probabilities below 0.5 for classifying the loan as default.

Figure 8a shows a loan with a low BERT score (0.37) and a SHAP value of −0.47, the highest in absolute value for that instance. This SHAP value counteracts the influence of other variables, leading to a (correct) classification of the loan as non-default. Conversely, Figure 8b illustrates the opposite scenario: a loan with a high BERT score (0.686), where the SHAP value—again, the highest for that instance—overrides the effect of other variables and correctly assigns the loan to the default class. Below are the description of both loans given by its borrowers:



(a) Loan (correctly) classified as non-default. (b) Loan (correctly) classified as default.

Figure 8: Waterfall plots representing the SHAP values of the features in two loans.

Purpose: Home improvement; BERT score: 0.37
"I want to have an in ground pool built with an aluminum screened enclosure. Nothing super fancy, just enough to enjoy with my wife a kids."

Purpose: Debt consolidation; BERT score: 0.686
"I have problems with one of my loans what I have it is with household-beneficial finance bank it is for $15,250.00 and the interest rate is 25.8% and the payment is 323 monthly and I have 3 months pass due, I page all time in time but the situacion make loss that way cause the gas bill was very higher and the last mont the second day my Mother die and was very expensive for us her funeral cause she die here but we bring her body to Nicaragua, and the rest is for a loan with American General Finance $2,500.00, Wells Fargo $480.00 , Fifth Third Bank Optimun $357.00, all make a total of $18,587.00, please I need cause I get some problems with the big loan with household-beneficial."

Although Section 6.2 suggested that shorter texts tend to have higher BERT scores, this is not the case here. BERT's ability to analyze content and textual characteristics enables it to correctly associate the long text with a high probability of default and the shorter one with a higher likelihood of repayment.

7.4 Impact of the BERT score in the classification results across the purpose categories

In this section, we examine how incorporating the BERT score in the model changed classification outcomes across various categories of the purpose variable. Given the language comprehension capabilities of LLMs, it is conceivable that the BERT score provides a more nuanced characterization of the risk associated to the loan purposes than the categorical *purpose* variable alone. For instance, the inherently ambiguous 'other' category may benefit from the nuanced understanding of loan descriptions provided by the BERT model, potentially leading to improved prediction outcomes.

Figure 9 illustrates the relative changes in balanced accuracy for each loan purpose after adding the BERT score. Notably, while the 'other' category shows a modest improvement of 2.11%, several other categories exhibit more significant enhancements, including 'educational' (9.22%), 'moving' (5.84%), 'medical' (3.84%), and 'small business' (3.40%). Given the black-box nature of our model, it is not easy to ascertain why these categories have improved more than the others. However, we posit that these categories share a commonality—the more detailed specification of purpose or a deeper understanding of the borrower's situation contributes to a more precise delineation of the default risk. For instance, 'education' loans might incorporate information about the borrower's field of study or educational institution, which could correlate with employability and repayment capacity. Similarly, in categories like 'moving', 'medical', or 'small business' loans, the BERT score likely reflects a deeper understanding of the borrower's situation, enabling a more accurate default risk assessment.

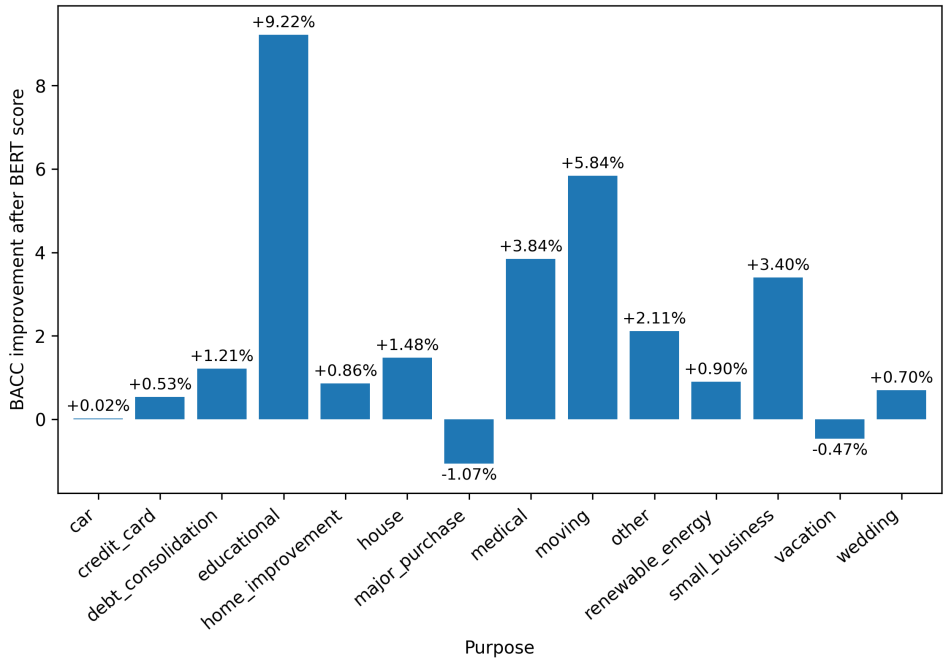


Figure 9: BACC changes (in relative percentage) by purpose after including the BERT score.

The following examples illustrate instances of loans that were initially misclassified as defaults without the BERT score but were correctly predicted as non-defaults after its inclusion:

Purpose: Educational; BERT score: 0.3765
"I'm 25 years old and living in New Orleans. I'm asking for a relatively small amount of money to help me take care of my post-Bac tuition for teacher certification and to help me pay off a credit card. I currently work as a private school teacher making very little money with no benefits (about $29,000 a year). I have to pay about $1000 in the coming year for my tuition, and I have to get health insurance ASAP, but it's hard to do so with no financial help from anyone else. My parents can't help me because my mother is permanently disabled and my father took a huge pay cut this year."

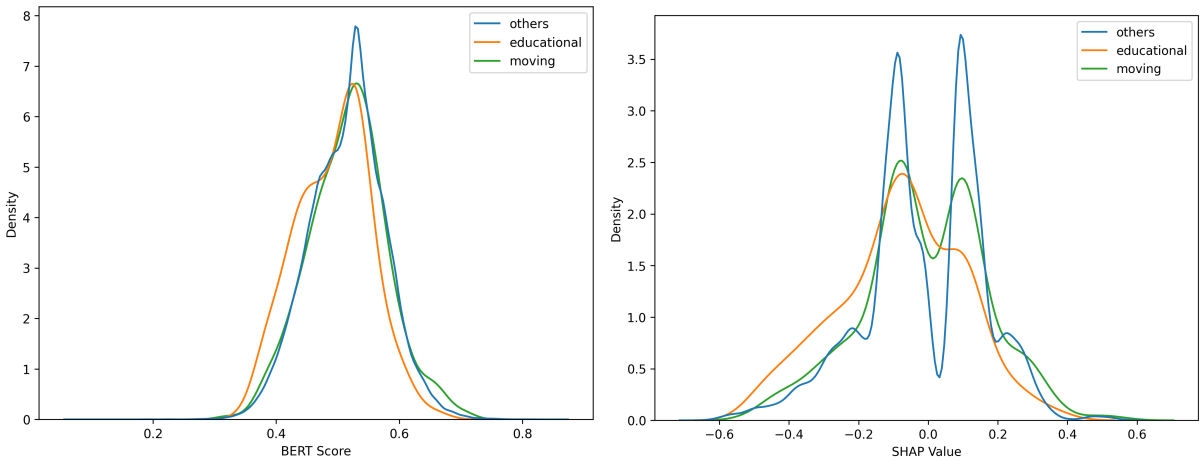
Purpose: Moving; BERT score: 0.3843
"Although I can afford payments,due to some recent expenses, I am short on cash flow for an unexpected move. I am, however, looking for a more reasonable alternative to banking rates. I have borrowed from Lending Club before and always paid fully and on time with automatic payments."

Purpose: Small business; BERT score: 0.3292
"The purpose of this loan is to fund advertising costs for a growing internet business venture. I am a successful sales professional earning an average of over $250K per year over the last 5 years. My credit scores are strong and I have a documented history of paying all my debts (personal or business related) on time."

Purpose: Medical; BERT score: 0.3921
"This loan will be used to pay off a Care Credit credit card currently at 21.9% I used the card to pay for a prosthetic limb that my health insurer would not cover."

In these cases, the corresponding BERT scores are consistently below 0.4. As shown in Figure 7, scores below this threshold are strong drivers for predicting non-default. All of these texts offer precise descriptions of the purpose of the loan or the borrower's situation, which allows a relatively moderate risk to be anticipated.

To better understand the differences in classifier performance across purposes, we conduct an additional analysis focusing on the purposes with the greatest performance improvement. Figure 10a shows



(a) Distributions of the BERT score.

(b) Distributions of the SHAP values.

Figure 10: BERT score and SHAP values distributions for the educational (orange), moving (green), and the rest of purposes (blue).

Table 14: Coefficients of the quantile regression at the 5th, 10th, 50th, 90th, and 95th percentiles.

	5th	10th	50th	90th	95th
Intercept	0.4101	0.4328	0.5178	0.5890	0.6106
Default	+0.0164*	+0.0166*	+0.0127*	+0.0124*	+0.0146*
Educational	−0.0208*	−0.0240*	−0.0205*	−0.0308*	−0.0316*
Moving	−0.0022	−0.0014	+0.0002	−0.0027	+2.8e-05
Default–Educational	−0.0067	−0.0077	+0.0126	+0.0288	+0.0084
Default–Moving	−0.0085	−0.0040	+0.0062	+0.0286*	+0.0357*

* p-value less than 0.01.

the distributions of BERT score values for loans with *educational* and *moving* purposes, compared to the distribution of loans with all other purposes, which serves as the reference. We can see that the reference distribution is more sharply peaked, indicating that most loans have average BERT score values that contribute little information to the classification. In contrast, the distribution of the loans with an *educational* purpose is shifted to the left and exhibits a higher density of low BERT score values (below 0.4), which indicates that the loan will likely be repaid. In the case of the *moving* purpose, the most pronounced difference can be seen in the right tail, as it has a higher frequency of values above 0.6, which helps to detect loans likely to default.

To assess the significance of these differences, we perform a quantile regression with interactions at five percentiles: the median (50th) and the two tails of the distribution (5th, 10th, 90th and 95th). The quantile regression model is specified as:

$$\text{BERT\_score}_i = \beta_0 + \beta_1 \cdot \text{Default}_i + \beta_2 \cdot D_E + \beta_3 \cdot D_M + \beta_4 \cdot (\text{Default}_i \times D_E) + \beta_5 \cdot (\text{Default}_i \times D_M) + \epsilon_i \quad (1)$$

Here, D_E and D_M represent the dummy variables for *educational* and *moving* loans, respectively, with all other purposes serving as the reference group. Meanwhile, Default_i is an indicator variable that denotes whether the loan eventually defaulted or not. The quantile regression allows us to assess whether there are significant differences in the BERT score between defaulted and non-defaulted moving and educational loans compared to loans of other purposes.

The results are presented in Table 14. At the 5th, 10th, 50th, 90th and 95th percentiles, the intercept values (0.4101, 0.4328, 0.5178, 0.5890, and 0.6106) capture the baseline BERT score for loans that do not belong to the educational or moving purposes and that were returned (i.e., $D_E = 0$, $D_M = 0$, and $\text{Default}_i = 0$). The positive and statistically significant coefficients for *default* across all quantiles

indicate that default is associated with an increase in BERT score (+0.0164, +0.0166, +0.0127, +0.0124, and +0.0146). Thus, the first conclusion that we can draw is that our fine-tuned BERT successfully assigns a higher risk to the loans that eventually defaulted, at least in the quantiles analyzed, in line with the results obtained in previous analyses.

The *educational* dummy variable shows significant negative coefficients across all quantiles (−0.0208, −0.0240, −0.0205, −0.0308, and −0.0316), confirming that, regardless of default status, educational loans have lower BERT scores at the quantiles analyzed compared to the reference group—consistent with a lower risk profile observed in Figure 10a. Regarding the interaction term, the *Default–Educational* coefficient is not significant at any quantile, suggesting that defaulted educational loans do not have significantly different BERT scores compared to repaid educational loans.

In contrast, the coefficients of the *moving* dummy when the *default* variable is zero are not statistically significant at any quantile. This implies that, in such cases, non-defaulted moving loans do not differ from the non-defaulted reference group in their baseline BERT scores. As for variable interaction, the *Default–Moving* coefficient is significant at the 90th and 95th percentiles (+0.0286 and +0.0357). This result indicates that, at the upper end of the BERT score distribution, defaulted loans from the moving purpose have higher BERT scores than defaulted loans from the rest of the purposes. In other words, the difference observed in Figure 10a in the right tail of the BERT score distribution for moving loans relative to the reference group is driven by the higher risk associated with defaulted loans.

Overall, these findings evidence a nuanced relationship: default increases BERT scores across the board, non-defaulted educational loans exhibit lower scores—reflecting lower risk—than other categories, and moving loans that eventually defaulted exhibit a higher risk reflected only in the upper quantiles. To conclude the analysis, Figure 10b shows the distribution of SHAP values for *educational* and *moving* purposes compared to the rest of the purposes. Again, educational loans show a distribution shifted to the left and with a denser left tail, indicating that XGBoost uses this variable to identify loans likely to be repaid. In the case of moving loans, differences are observed at the extremes, especially in the right tail, indicating the improvement in discrimination that this variable provides to the classifier. This analysis shows that the interaction between the BERT score and the educational and moving purposes is the reason behind the remarkable performance improvements observed in these categories.

As a conclusion, while previous research had reported mixed results regarding the predictive power of linguistic factors for loan default [16, 54, 15, 46], our findings suggest that an LLM-based risk indicator used at the granting stage has a significant positive influence in the classification results, thus demonstrating BERT's capacity to draw meaningful information from the loan descriptions.

8 Conclusion

In this paper, we presented a novel approach that leverages state-of-the-art natural language processing techniques to enhance credit risk models. By fine-tuning BERT on loan descriptions, we generated a risk score that effectively distinguishes between defaulted and non-defaulted loans, particularly at the granting stage, when decision-makers have limited variables available to inform their decisions. In addition, integrating this BERT-based risk score with traditional variables significantly improved the performance of conventional loan-granting models. This result aligns with those obtained by Xia et al. [57]. Our analysis suggests that the information extracted by the language model can capture aspects of the text related to the linguistic aspects but also with content-based factors related to loan purpose and the borrower's creditworthiness. The inclusion of the BERT-based risk score also reshapes the classifier's decision-making process and the role played by other variables.

Our approach can be easily applied without the need for manual annotation, which is a complex and subjective task. Additionally, while fine-tuning the LLM is a computationally intensive process that requires GPU resources, generating predictions such as the risk score for a loan description can be done rapidly on standard home equipment, making this approach highly accessible. This work opens several avenues for further exploration to refine both the predictive capability of the model and our understanding of loan applicants' situations.

In our work, we have thoroughly documented the model's architecture as well as the preprocessing, fine-tuning, and training procedures. We have also used SHAP values to understand how the final model classifies and even how it makes decisions for individual instances. While these aspects are crucial in

financial applications, a key limitation remains: the lack of transparency in how the BERT-based risk score is generated, which limits the understanding of the factors influencing the score and its potential biases. This challenge hinders the practical application of this approach in real-world settings, where it is critical to understand the score generation process and ensure that it does not introduce biases related to gender, ethnicity, or financial inclusion. Therefore, enhancing the transparency and explainability of the BERT score is essential, not only for regulatory compliance but also to build trust among borrowers and lenders [43].

In this respect, several approaches exist to better understand how neural-based Natural Language Processing models such as BERT work [31]. Intrinsic methods, such as inspecting attention weights, do not offer sufficient transparency or meaningful explanations for model predictions [21, 45]. Still, it is possible to use surrogate models, such as LIME [41] or *anchor* rules [42], to try to interpret BERT-based predictions, as has been done in fake news detection [51]. Such an approach can help to better understand how these models work and facilitate their adoption in real-world settings.

A different approach that could also be explored is the use of LLM-based topic modeling techniques, such as BERTopic [18]. It could help to identify "risky topics" by making use of the deep semantic understanding of LLMs. These topics could then be used as additional input variables in the granting model, improving both predictive performance and transparency in the decision-making process. Additionally, the embeddings from BERT or other LLMs could be used to generate interpretable topics from large, complex vocabularies, as demonstrated in other studies [14]. Applying such topic modeling to loan descriptions could help identify loans with varying risk levels, further enhancing the transparency of credit risk assessments.

Further research could also explore the use of advanced LLMs. Encoder-only models like RoBERTa [28] may capture more intricate linguistic patterns while emerging generative AI approaches could redefine how risk scoring is performed. Although these models were not originally designed for such tasks, recent advancements have shown promising approximations. For instance, techniques like CARP (Clue And Reasoning Prompting) [49] utilize in-context learning with few-shot examples to perform classification without fine-tuning. Applying such methods to risk scoring may open new possibilities for achieving robust results while minimizing computational overhead.

Future work should also address the economic implications of our findings by integrating cost- or profit-sensitive approaches [56, 3]. Investigating how the inclusion of textual descriptions impacts financial outcomes could provide valuable insights into the practical utility of these methods. By linking improved prediction performance to tangible economic benefits, we can further bridge the gap between academic innovation and real-world application.

Acknowledgements

We thank Antonio Caparrini López and Miller Janny Ariza Garzón for their help at different stages of the project.

References

[1] Shabeen A. Basha, Mohammed M. Elgammal, and Bana M. Abuzayed. Online peer-to-peer lending: A review of the literature. *Electronic Commerce Research and Applications*, 48:101069, 2021.

[2] Miller Janny Ariza-Garzón, Javier Arroyo, Antonio Caparrini, and Maria-Jesus Segovia-Vargas. Explainability of a machine learning granting scoring model in peer-to-peer lending. *IEEE Access*, 8:64873–64890, 2020.

[3] Miller-Janny Ariza-Garzón, Javier Arroyo, María-Jesús Segovia-Vargas, and Antonio Caparrini. Profit-sensitive machine learning classification with explanations in credit risk: The case of small businesses in peer-to-peer lending. *Electronic Commerce Research and Applications*, 67:101428, 2024.

[4] Miller-Janny Ariza-Garzón, María-Del-Mar Camacho-Miñano, María-Jesús Segovia-Vargas, and Javier Arroyo. Risk-return modelling in the p2p lending market: Trends, gaps, recommendations and future directions. *Electronic Commerce Research and Applications*, 49:101079, 2021.

[5] Miller Janny Ariza-Garzón, Mario Sanz-Guerrero, Javier Arroyo Gallardo, and Lending Club. Lending Club loan dataset for granting models, May 2024. https://doi.org/10.5281/zenodo.11295916.

[6] Adithya Bhaskar, Alexander R Fabbri, and Greg Durrett. Prompted opinion summarization with GPT-3.5. *arXiv:2211.15914*, 2022.

[7] Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv:2005.14165*, 2020.

[9] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *Practical ML for Developing Countries Workshop at ICLR 2020*, 2020.

[10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Greg Brockman, Alex Ray, et al. Evaluating large language models trained on code. *arXiv:2107.03374*, 2021.

[11] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[12] Mark Cummins, Theo Lynn, Ciarán Mac an Bhaird, and Pierangelo Rosati. *Addressing Information Asymmetries in Online Peer-to-Peer Lending. In Disrupting Finance*, page 15–31. Springer International Publishing, December 2018.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 07 2020.

[15] Gregor Dorfleitner, Christopher Priberny, Stephanie Schuster, Johannes Stoiber, Martina Weber, Ivan de Castro, and Julia Kammler. Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms. *Journal of Banking & Finance*, 64:169–187, 2016.

[16] Qiang Gao and Mingfeng Lin. Lemon or cherry? The value of texts in debt crowdfunding. Technical Report 18, Center for Analytical Finance. University of California, Santa Cruz, 2015.

[17] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. Target-dependent sentiment classification with BERT. *IEEE Access*, 7:154290–154299, 2019.

[18] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv:2203.05794*, 2022.

[19] Michal Herzenstein, Scott Sonenshein, and Utpal M. Dholakia. Tell me a good story and I may lend you my money: The role of narratives in peer-to-peer lending decisions. *SSRN Electronic Journal*, 2011.

[20] John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, Cambridge, Massachusetts, USA, 04 1992.

[21] Sarthak Jain and Byron C. Wallace. Attention is not explanation. *CoRR*, abs/1902.10186, 2019.

[22] Cuiqing Jiang, Zhao Wang, Ruiya Wang, and Yong Ding. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, 266(1–2):511–529, October 2017.

[23] Johannes Kriebel and Lennart Stitz. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, 302(1):309–323, October 2022.

[24] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942*, 2019.

[25] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[26] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv:1910.13461*, 2019.

[27] Yuelei Li, Aiting Hao, Xiaotao Zhang, and Xiong Xiong. Network topology and systemic risk in peer-to-peer lending market. *Physica A: Statistical Mechanics and its Applications*, 508:118–130, 2018.

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019.

[29] Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.

[30] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[31] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys*, 55(8), December 2022.

[32] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.

[33] Jeremy Michels. Do unverifiable disclosures matter? Evidence from peer-to-peer lending. *The Accounting Review*, 87(4):1385–1413, 2012.

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.

[35] André Aoun Montevechi, Rafael de Carvalho Miranda, André Luiz Medeiros, and José Arnaldo Barra Montevechi. Advancing credit risk modelling with machine learning: A comprehensive review of the state-of-the-art. *Engineering Applications of Artificial Intelligence*, 137:109082, 2024.

[36] David Pride, Matteo Cancellieri, and Petr Knoth. CORE-GPT: Combining open access research and large language models for credible, trustworthy question answering. *arXiv:2307.04683*, 2023.

[37] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6639–6649, Red Hook, NY, USA, 2018. Curran Associates Inc.

[38] Zhiyuan Qi, Dongyu Chen, and Jennifer J. Xu. Do facial images matter? Understanding the role of private information disclosure in crowdfunding markets. *Electronic Commerce Research and Applications*, 54(C), jul 2022.

[39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1), jan 2020.

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: high-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[43] ROFIEG. Thirty recommendations on regulation, innovation and finance. final report to the european commission by the expert group on regulatory obstacles to financial innovation. Technical report, European Commission, 2019.

[44] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108*, 2019.

[45] Sofia Serrano and Noah A. Smith. Is attention interpretable? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics.

[46] Michael Siering. Peer-to-peer (p2p) lending risk management: Assessing credit risk on social lending platforms using textual factors. *ACM Transactions on Management Information Systems*, 14(3), jun 2023.

[47] Matthew Stevenson, Christophe Mues, and Cristián Bravo. The value of text for small business default prediction: A Deep Learning approach. *European Journal of Operational Research*, 295(2):758–771, December 2021.

[48] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics*, pages 194–206, Cham, 2019. Springer International Publishing.

[49] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv:2305.08377*, 2023.

[50] Xu Sun and Weichao Xu. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.

[51] Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. New explainability method for BERT-based model in fake news detection. *Scientific Reports*, 11(1), Dec 2021.

[52] Vijay Srinivas Tida and Sonya Hy Hsu. Universal spam detection using transfer learning of BERT model. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, pages 7669–7677, 2022.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[54] Shuxia Wang, Yuwei Qi, Bin Fu, and Hongzhi Liu. Credit risk evaluation based on text analysis. *International Journal of Cognitive Informatics and Natural Intelligence*, 10:1–11, 01 2016.

[55] Yufei Xia, Lingyun He, Yinguo Li, Nana Liu, and Yanlin Ding. Predicting loan default in peer-to-peer lending using narrative data. *Journal of Forecasting*, 39(2):260–280, 2020.

[56] Yufei Xia, Chuanzhe Liu, and Nana Liu. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications*, 24:30–49, 2017.

[57] Yufei Xia, Zhengxu Shi, Xiaoying Du, and Qiong Zheng. Extracting narrative data via large language models for loan default prediction: when talk isn't cheap. *Applied Economics Letters*, page 1–6, November 2023.

[58] Jennifer Xu, Dongyu Chen, and Michael Chau. Identifying features for detecting fraudulent loan requests on p2p platforms. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 79–84, 2016.

[59] Jennifer Xu, Dongyu Chen, Michael Chau, Liting Li, and Haichao Zheng. Peer-to-peer loan fraud detection: Constructing features from transaction data. *MIS Quarterly*, 45(3):1777–1792, September 2022.

[60] Jianrong Yao, Jiarui Chen, June Wei, Yuangao Chen, and Shuiqing Yang. The relationship between soft information in loan titles and online peer-to-peer lending: evidence from renrendai platform. *Electronic Commerce Research*, 19(1):111–129, 2018.

[61] Xin Ye, Lu an Dong, and Da Ma. Loan evaluation in p2p lending based on random forest optimized by genetic algorithm with profit score. *Electronic Commerce Research and Applications*, 32:23–36, 2018.

[62] Weiguo Zhang, Chao Wang, Yue Zhang, and Junbo Wang. Credit risk evaluation model with textual features from loan descriptions for p2p lending. *Electronic Commerce Research and Applications*, 42:100989, 2020.