Social Capital as Soft Information in Predicting Microfinance Loan Repayment

Jesse E. Vlaswinkel
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT

Loan repayment success is a vital part of the continuation of microfinance institutions (MFIs), making the ability to predict loan repayment success an important aspect of loan decision-making. Using a large dataset from an MFI in the Netherlands, this research investigates whether borrowers' social capital can predict loan repayment success, utilizing a keyword-based method. Social capital is operationalized through bonding, bridging, and linking. Bonding (strong ties such as family and close friends) shows the strongest and most consistent predictive power; however, overall results are weak, indicating social capital alone is not a strong predictor of successful loan repayment. This contradicts prior literature that understates the economic role of bonding social capital. Furthermore, it indicates that not all types of soft information are strong predictors of loan repayment success. Limitations of this research stem from the social capital extraction method (keyword search) and the procedures used by loan officers to record soft information. This research encourages MFIs to explore new ways to document soft information. Nevertheless, this research highlights the contextual challenges and predictive potential of soft information in microfinance.

Graduation Committee members:
Franziska Koefer, MSc
Dr. Maximilian Goethner

Keywords
Microfinance, Soft Information, Social Capital, SME Lending, Loan Repayment, Europe

During the preparation of this work, the author used ChatGPT to assist in coding in Orange and to refine the text. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

1 INTRODUCTION

1.1 Situation and Complication

Loan assessment and repayment prediction are key activities for microfinance institutions. Screening leads to more successful repayment and more lucrative start-ups (Soomro et al., 2024). To assess potential customers, loan officers seek both hard and soft information about them. One method by which banking firms use hard information is through credit scores (Liberti & Petersen, 2018).

The use of primarily hard information, such as credit-score models, faces challenges when applied to decision-making in microfinance. Firstly, people applying for these smaller loans often do not have an elaborate financial history (Schreiner, 2004). Secondly, over-reliance on credit scores could lead to social and economic inequality (Kim et al., 2023). To mitigate this limitation, small banks rely more heavily on soft information. This allows small banks to overcome the problem of a limited credit history, and it also enables the creditor to take advantage of their proximity to the customers (McCann & McIndoe-Calder, 2015).

There is contradictory evidence on the use of soft information in loan decision-making. Research done with data from a Tunisian microfinance bank shows that making use of soft information, together with hard information, can have positive effects. In that particular study, the rates of misclassification of both false positives and false negatives were reduced (Baklouti & Bouri, 2014). In contrast, a study from Campbell et al. (2019) revealed that the use of soft information can in some cases also lead to worse loan decisions, mostly due to unwanted factors. An example highlighted poorer loan outcomes when the loan intake discussion was scheduled just before the weekend or when both the officer and the applicant were of the same gender.

1.2 Research Objective

Generally, literature states that the use of soft information, exclusively or in conjunction with hard information, is still relevant and is even preferred in small bank loan decision-making (Y. Chen et al., 2015; McCann & McIndoe-Calder, 2015). However, the subjectivity of this information has the potential to negatively affect loan quality. Although there are suggestions on how to make soft information gathering more insightful and objective, less research has been done on which kind of soft information predicts loan repayment more accurately. If predictions are improved, the loan decision process has the potential to be less costly and more reliable (Cornée, 2019). At the same time, there is still a need for research to determine to what degree human judgment remains relevant when making decisions using soft information and whether current content analysis methods converge with human loan decisions.

This paper seeks to find out how significant the impact of soft information is on loan repayment performance. Specifically, it aims to find out if social capital is a strong predictor of loan repayment success. The following question is formulated:

'To what extent can borrowers' social capital -subdivided into bonding, bridging, and linking- predict loan repayment performance in microfinance during the loan screening process?'

1.3 Contributions

One of the main academic contributions of this research is its examination of the use of soft information in a developed economy, as most previous research has focused on developing countries. Additionally, this research will also utilize a larger dataset than those used in prior studies. Furthermore, the study will focus specifically on individual loan repayment success, rather than group repayment success. Lastly, this work adds to the understanding of the application of soft information in loan decision-making, particularly within the frameworks of bonding, bridging, and linking social capital, as well as screening and signalling theory.

With regard to its practical contribution, this paper aims to improve the loan evaluation process, thereby increasing the efficiency of the capital market. In addition, it has the potential to enhance the accuracy of loan repayment predictions, thereby enabling microcredit institutions to make more informed loan decisions. Moreover, improved loan outcome success has the potential to also increase (start-up) business success. Lastly, by advocating for the use of soft information alongside hard information, it could help promote greater equality while ensuring fairness.

2 LITERATURE REVIEW

2.1 Soft Information

Soft information is qualitative, contextual information that cannot be credibly transferred to others. Hard information is quantitative and impersonal information that is independent of the collection process (Petersen, 2004). Rather than viewing hard information and soft information as opposites, they could instead be viewed as a continuum (Liberti & Petersen, 2018).

Financial institutions can gather soft information through loan discussions, visits to the borrowing enterprise, and interviews with employees (Y. Chen et al., 2015). Larger banks prefer to use credit scores instead of soft information for loan decision-making, and thus maintain a more impersonal relationship with their customers (Berger et al., 2005). This approach is reflected in their preference for larger customers, as smaller customers often lack sufficient credit history and adequate collateral.

Lack of access to credit is the most common issue faced by micro- and small entrepreneurs, largely due to their exclusion from large credit facilities (Nawai & Shariff, 2010). To overcome this obstacle, small firms invest in creating information that supports them in accessing credit. For example, the legal form of the business can play a significant role: enterprises with high paid-in capital have disproportionately greater access to credit compared to those with low paid-in capital legal structures (Bracht et al., 2024).

2.2 Loan Repayment

The outcome of a loan can be classified in multiple ways. A loan may be repaid on time without any issues, or it may experience repayment delays if a payment is late. In cases where delays continue to persist and the likelihood of recovery becomes minimal, the loan is classified as defaulted (Addae-Korankye, 2014). Typically, microfinance institutions (MFI) consider loans successful when the loan is fully repaid on time, or within the negotiated grace period. Loan repayment issues are a major reason for the failure of MFIs (Kiros, 2020).

Defaulting on a loan or missing a payment negatively affects both the lender and the borrower. The lender forgoes interest income, suffers opportunity costs on the principal, and incurs legal costs and other related expenses. On the other hand, the borrower must decide whether to default and face the consequences, or continue making repayments despite financial difficulties, in order to preserve their reputation and avoid the costs associated with defaulting (Ntiamoah et al., 2014).

A key factor in both loan repayment success and entrepreneurial success is loan negotiation, during which the loan amount, the interest rate, and the collateral are discussed and determined. Borrowers with strong soft information tend to pay lower interest rates, require less collateral, and have access to more credit (Y. Chen et al., 2015).

2.3 Theoretical Framework

To better understand the behavior of the creditor and debtor, signalling and screening theory helps to explain how both parties communicate and interpret information.

Signalling theory can be used to describe the behavior of two parties when there is an information asymmetry (Connelly et al., 2011). Signalling theory explains how and to what extent the sender communicates information, and how the receiver interprets the signal and what decision is made. The application of signalling theory to lending shows that high-quality potential customers display their invisible qualities by taking part in visible activities that are expensive and hard for low-quality customers to copy. The signalling theory has three core concepts: signals are alterable, signalling costs have a negative relation with actors' quality, and signallers perform better than non-signallers (Bafera & Kleinert, 2023).

In comparison, the screening theory explains how the lender searches and uses external cues to surmount information asymmetry (Zhang et al., 2024). Thorough screening leads to more successful loan outcomes and lower collateral requirements (Manove et al., 2001).

Both signalling and screening theory are applied during the loan decision-making process. For the purposes of this research, the signalling and screening of soft information is particularly important. The lender primarily benefits from reduced information asymmetry, while creditworthy borrowers also benefit from proper identification through signalling and screening. The creditor will try to seek out soft information cues that indicate a successful loan repayment, whilst a creditworthy debtor will emit strong soft information signals to gain access to credit, lower interest rates, and reduce collateral requirements.

The soft information signalled and screened contribute to improved loan prediction accuracy. Certain aspects of soft information serve as stronger predictors than others. Research based on data from a Taiwanese finance company showed that leadership, customer relationships, public praise, and team quality are significant predictors of loan default (Y. Chen et al., 2015).

Another study, examining loan repayment performance in microfinance institutions in Pakistan, identified business performance as the strongest predictor. However, business performance was strongly influenced by the borrower's social capital, and social capital also has a modest direct effect on repayment performance (Iqbal & Rao, 2023). In this particular research, social capital was defined as the relationship between the borrower and the lending institution, and was operationalized through aspects such as the frequency of communication and participation in workshops organized by the MFI. The terms of the loan were also found to have an impact, although minor, on both business performance and repayment outcomes.

Social capital is a broad term with multiple definitions and ways of measurements. One definition states that it consists of three families: trust, ease of cooperation, and networks (Paldam, 2000). A further division of social capital comes in the form of the network approach, which differentiates between bonding, bridging, and linking social capital.

2.3.1 Bonding

In bonding social capital, emotional support is provided by strong ties, e.g. close friends and family (Nguyen et al., 2013). Bonding provides links between homogeneous people, whilst excluding non-members. Different researchers suggest different effects from bonding social capital. Some propose that practices such as nepotism and clientelism cause benefits for the ingroup, but overall negative effects on society; even stating that bonding social capital has no effect or a negative effect on economic performance (Claridge, 2018). In contrast, others state that bonding provides socio-economic development with its social control and supporting nature. A third view is that the effect of bonding social capital is largely context dependent (Muringani et al., 2021).

2.3.2 Bridging

On the other hand, bridging social capital is not provided by strong ties, but instead by connections with people or organizations such as colleagues or associations. It provides links between heterogeneous people in open networks (Nguyen et al., 2013). Bridging social capital has been associated with improvements in economic development, growth, and employment (Claridge, 2018). Furthermore, due to the heterogeneous connections which increase knowledge spread,

bridging social capital is associated with innovation, firm entry, and entrepreneurship (Muringani et al., 2021).

2.3.3 Linking

More recently, a third type was introduced: linking social capital, which refers to the relationships between ordinary citizens and individuals in positions of power (Aldrich & Meyer, 2015). The strength of linking social capital is determined by the position of the person with whom the link is, the higher the rank, the more powerful the link. Linking social capital is susceptible to a patron–client relationship, which can be exploitative (Dufhues et al., 2011).

Bonding, bridging, and linking social capital are not mutually exclusive and can be present at the same time; in fact, they often occur together. High bonding and high bridging social capital being present at the same time produces positive socio-economic outcomes. While high bridging and low bonding social capital could suffer from a lack of sanctions to follow up on common expectations. Low bridging and high bonding result in nepotism and individualism (Muringani et al., 2021). Research done by Dufhues et al. (2011) on the effect of the three different types of social capital on loan repayment in two countries in Southeast Asia using a survey, found that only bonding has a positive effect on loan repayment in Thailand. In comparison, in Vietnam, only bridging in combination with linking was identified to have a significant positive effect.

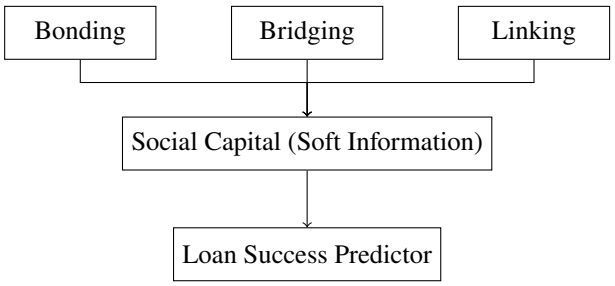


Figure 1: Conceptual framework of loan success predictors based on soft information

2.3.4 Hypothesis

Based on the literature and the previously discussed dynamic between signalling and screening theory, the following hypothesis is proposed to address the research question:

H: *Social capital, specifically bridging and linking, has a positive impact on the ability of microfinance institutions to predict loan repayment success.*

3 METHODOLOGY

3.1 Research Design

This paper seeks to investigate the impact that social capital has on loan success. As the data includes both numerical (binary) variables and unstructured textual information, quantitative research is applied. This method enables the analysis to assess statistical patterns, whilst using natural language processing to identify the social capital in the textual information.

Natural language processing (NLP) refers to a technique that allows machines to process and analyze human language (Cambria & White, 2014). Although recent advances in artificial intelligence (AI) and large language models have led to more advanced and sophisticated NLP approaches, this study applies a more conventional, rule-based method of NLP (Jonker et al., 2024). To be more specific, keyword matching was used to seek the presence of social capital indicators in the soft information. This method enabled the transformation of unstructured text into binary values for use in the analysis.

Most NLP methods are either 'rules-based', where the person is in full control of how the program works. Or they are 'machine-learning', where a training data set is given to the computer, and then it uses a large data set to deliver the answer (Coulthard & Taylor, 2022). Given the size of the dataset used in this research, using content analysis, manually assigning values to the records would be infeasible in the time frame of this research. However, some level of human input remains necessary to define, classify, filter, and interpret the data or documents. Therefore, this study applied a hybrid approach that combines the transparency of rule-based keyword matching with the efficiency and automation offered by NLP techniques.

3.2 Dataset and Feature Description

The dataset used in this research consists of more than 14,000 loan application records from 2018 to 2021, sourced from a microfinancing institution in the Netherlands.

The records contain hard information such as the loan amount, interest, and age of the borrower. In addition, the dataset includes soft information in Dutch about the borrower's business activities, entrepreneur background, target market, clarifications regarding financial analysis and private circumstances of the borrower, as well as the conclusion of the risk manager and loan officer. This soft information was written down by different employees. Since each of these columns could potentially include social capital information, all columns were combined into a new variable called *'Alltext'*. The records include binary indicators for each repayment outcome, specifying whether the borrower defaulted, repaid with or without issues, or if there was a deferment of repayment either due to COVID-19 or due to general reasons.

In this research, only disbursed loans and loans with a known outcome were used, excluding those still in progress. This was done because these loans would otherwise be marked as unsuccessful even though their outcome is still unknown.

Loan repayment success was defined solely as the column *Loan repaid fully without issues*, excluding the column *Loan repaid (fully or partially) with repayment issues*. This criterion was chosen because the column *Loan repaid (fully or partially) with repayment issues* does not state to what extent the different loans were repaid. Furthermore, it is also unknown if these loans were repaid within the negotiated

grace period.

3.3 Data Analysis

Initially, to detect whether social capital was present in the *'Alltext'* column, using a supervised machine learning method, 200 rows were manually classified for the presence of social capital as a training dataset. However, this method quickly proved flawed as it was too subjective and the results were not accurate. To achieve a more effective method, a new list of keywords (or dictionary) was developed based on literature (Appendix 10.1). This method was chosen because it is least biased and most replicable; furthermore, research shows that it is better and more accurate than costlier methods such as query expansion techniques and topic model-based classification rules. Only active supervised learning shows stronger performance, but this comes at the cost of reduced replicability (Wankmüller, 2023). A dictionary or keyword method has particularly strong interpretability and is especially useful when the construct has a sufficient theoretical base (Herhausen et al., 2025). These keywords are divided into three recognized types of social capital: the aforementioned bonding, bridging, and linking. The dataset was then fed into the program Orange Data Mining, which is built on Python (Appendix 10.2).

Text preprocessing is not used in this research for several reasons. First of all, by using the Python script, it is not necessary to turn the raw text into a Corpus object. Additionally, the integral text preprocessing applications of Orange, or stemming methods based upon English, are not specifically built for Dutch, making them prone to mistakes and unusable normalization (Gaustad & Bouma, 2001). Furthermore, settings like document frequency are not relevant, as only the presence of a word matters, not how often it appears. Finally, without text preprocessing, the meaning of words remains more relevant and interpretable (Chai, 2022).

Sequentially, a Python script was written to classify whether different types of social capital were present in the column *'Alltext'* using a keywords search (Appendix 10.3).

Two analyses with the dataset were done:

1. **Analysis 1:** Including loans fully repaid without issues, and loans with repayment delays or issues excluding those with deferments due to COVID-19
2. **Analysis 2:** Including loans fully repaid without issues, and loans with repayment delays or issues excluding those with deferments (both general reasons and COVID-19)

The decision to create two separate analyses was based upon two reasons. Firstly, both analyses have a different class imbalance between successful repayments and unsuccessful repayments, which is particularly apparent in chart 2. Due to the class imbalance, models like random forest have a tendency to simply predict the majority outcome class each time, leading to high accuracy but lower minority class prediction (C. Chen et al., 2004). To combat this, studies recommend reducing the class imbalance in the training sets (Dube & Verster, 2024). To achieve this, in both models the 'balance class distribution' option was activated, ensuring each fold has about as many successful as unsuccessful repayment records. The second reason to conduct two separate analyses is to examine the effect the presence of social capital has on repayment success when including deferred loans, as these loans could be considered to be not successfully repaid.

The following step was to compute whether the different types of social capital had an impact on the success of loan repayment. To calculate the relation between social capital and loan repayment success, firstly a logistic regression was used. A logistic regression is a statistical model that predicts the probability of a binary outcome based on one or multiple variables (Peng et al., 2002). Secondly, a random forest model was used to search for strong predictive performance and to ensure no relationship was missed. A random forest creates multiple decision trees based on its given data and variables, and consolidates these outcomes into a single prediction (Breiman, 2001a). This approach was chosen in accordance with Breiman (2001b), who supports the use of algorithmic models such as random forest to try and uncover patterns statistical models might fail to detect.

To evaluate the predictive power of the logistic regression and random forest, 10-fold stratified cross-validation was applied using Orange's 'Test and Score'. This ensured a proper class distribution in each test fold and reduced randomness. Control variables were intentionally not used in this research, as in addition to investigating the relation between social capital and loan repayment success, another aim of this study was to review to what extent computer models are able to identify social capital in unstructured text. For this reason, introducing structured hard information control variables such as interest rate or loan amount could overpower the weak but possibly meaningful predictive power of the extracted soft information.

Finally, to test whether individual keywords had significant predictive power over repayment success, a separate random forest was run for each keyword. This was done solely on the dataset excluding the deferments of repayment due to general reasons (analysis 2).

4 RESULTS

4.1 Keyword Search Outcome

After filtering the loans in the dataset to include only those that were disbursed, finished, and not deferred -either due to COVID-19 or due to general reasons- the keyword search found the following number of entries for each type of social capital. It is important to note that individual records can contain multiple types of social capital.

Category	Analysis 1	Analysis 2
Bonding	5,459	2,766
Bridging	2,578	1,330
Linking	2,039	1,004
Total Rec.	7,316	3,693

Table 1: Number of entries with social capital per category

Bonding social capital was detected in the largest quantities, approximately 75% in both analyses. Bridging and linking social capital were present in around 36% and 28% of the cases, respectively.

In table 2, the most frequently found keywords are presented for analysis 2. The results for analysis 1 were similar for the bonding and bridging categories. However, in the linking category, employer and school switched places, and government and official also changed places.

Bonding	Number	Bridging	Number	Linking	Number
Partner	3,725	Network	1,742	Municipality	601
Children	1,662	Acquaintance	310	Employer	553
Father	827	Community	309	School	312
Mother	649	Colleague	246	Official	286
Family	610	Abroad	201	Government	236

Table 2: Most found keywords analysis 2. (Translated to English)

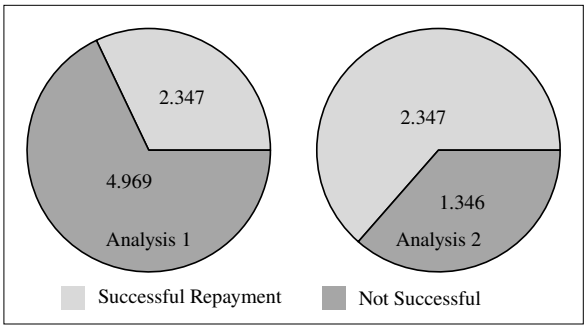


Figure 2: Distribution of repayment outcomes in filtered datasets

In the keyword search that excludes all deferments, the percentage of successful repayments is significantly higher, reflecting the exclusion of loans impacted by external circumstances such as COVID-19 or other general deferment reasons.

4.2 Predictive Modeling Results

To analyze the performance of the logistic regression and random forest, four evaluation metrics are used. The area under ROC curve (AUC) measures the extent to which the model can distinguish between borrowers who successfully repaid and those who did not. F1-score balances the mean of borrowers who repay (recall) with the accuracy of these predictions (precision), by taking into account both false positives and false negatives. Due to the class imbalance evident in figure 2, F1 is a valuable metric, as it penalizes extreme values either way. On the other hand, precision displays what portion of borrowers predicted to repay actually did. Finally, the Matthews Correlation Coefficient (MCC) takes into account all possible outcomes in the confusion matrix. Resulting in a single score -with 1 showing perfect prediction, 0 random guessing, and -1 a completely inverted relation between prediction and observations- that is especially relevant even when classes are unbalanced (Hicks et al., 2022).

Type & model	AUC	F1	Prec	MCC
Bonding				
Logistic Regression	0.505	0.406	0.580	0.025
Random Forest	0.509	0.406	0.580	0.025
Bridging				
Logistic Regression	0.494	0.548	0.560	-0.009
Random Forest	0.502	0.436	0.567	-0.006
Linking				
Logistic Regression	0.478	0.532	0.551	-0.030
Random Forest	0.489	0.500	0.558	-0.013
Bonding & Bridging				
Logistic Regression	0.503	0.406	0.580	0.025
Random Forest	0.511	0.477	0.573	0.017
Bonding & Linking				
Logistic Regression	0.493	0.406	0.580	0.025
Random Forest	0.498	0.427	0.575	0.018
Bridging & Linking				
Logistic Regression	0.483	0.534	0.557	-0.017
Random Forest	0.507	0.528	0.567	0.007
Bonding & Bridging & Linking				
Logistic Regression	0.495	0.417	0.576	0.020
Random Forest	0.506	0.494	0.571	0.013

Table 3: Model Results Analysis 1

When reviewing the performance of different types of social capital in the analysis that excludes COVID-19 deferments (analysis 1), bonding appears to be the strongest and most consistent performer. It achieved a small positive MCC and relatively high precision compared to the other types. The F1 score is noticeably lower, which—given the high precision—is due to low recall. This can be explained by the model being cautious and missing true positives. In contrast, bridging and linking, both individually and in combination, perform poorly. Overall, the random forest seems to slightly outperform the logistic regression

Type & model	AUC	F1	Prec	MCC
Bonding				
Logistic Regression	0.504	0.563	0.556	0.040
Random Forest	0.508	0.563	0.556	0.040
Bridging				
Logistic Regression	0.478	0.514	0.521	-0.034
Random Forest	0.475	0.471	0.521	-0.031
Linking				
Logistic Regression	0.498	0.429	0.551	0.023
Random Forest	0.506	0.429	0.551	0.023
Bonding & Bridging				
Logistic Regression	0.509	0.541	0.541	0.009
Random Forest	0.507	0.534	0.539	0.006
Bonding & Linking				
Logistic Regression	0.510	0.563	0.556	0.040
Random Forest	0.506	0.533	0.539	0.005
Bridging & Linking				
Logistic Regression	0.501	0.447	0.543	0.012
Random Forest	0.503	0.545	0.543	0.013
Bonding & Bridging & Linking				
Logistic Regression	0.513	0.542	0.546	0.020
Random Forest	0.514	0.548	0.551	0.030

Table 4: Model Results Analysis 2

In the analysis that filtered out all generally deferred loans (analysis 2), bonding once again stands out as the most effective and consistent predictor, achieving the highest MCC and precision of all combinations. Combining bonding with linking also preserved the high precision and MCC. Linking performed better individually as well, compared to the previous analysis. Bridging continues to underperform, even resulting in negative MCC values and the weakest AUC. The combination of all three types of social capital also showed promising results, especially with the random forest model. Random forest outperformed logistic regression once more, although the differences remain minimal.

4.3 Individual Keyword Strength

To gain insight into which individual keywords possess what predictive power, an individual keyword strength test was conducted. The evaluation was done using the dataset excluding the generally deferred records (analysis 2), using MCC from the random forest as this performed best in the previous analyses.

Bonding	**MCC**
Partner	0.057
Mother	0.048
(Girl)friend (f)	0.033
Children	0.030
Family	0.026
(Boy)friend (m)	0.019
Father	0.019
Child	0.013
Wife	0.009
Uncle	0.003
Grandfather	0.001
Grandmother	0.001
Stepmother	0.001
Neighbor (f)	0.001
(Life)partner	0.001
Stepfather	0
Aunt	0
Daughter	0
Neighbor (m)	0
Living together	0
Friends group	0
Niece (cousin)	-0.001
Neighbors	-0.001
Housemate (roommate)	-0.001
Informal caregiver	-0.001
Nephew (cousin)	-0.002
Son	-0.003
Church	-0.003
(Immediate) family	-0.003
Husband	-0.003

Bridging	**MCC**
Acquaintance	0.009
Network	0.008
Club	0.003
Charity work	0.002
Sports club	0.001
Networking meeting	0.001
Association	0
Multicultural	0
Abroad	0
Competition	0
Intercultural	0
Community center	0
Community	-0.001
Volunteer	-0.001
Foundation	-0.002
Culture	-0.004
Colleague	-0.008

Linking	**MCC**
Employer	0.029
Government official	0.013
Municipality	0.010
School	0.005
Boss	0.001
Manager	0.001
Sponsor	0.001
Government	0
Civil servant	0
Court	0
Ministry	0
Mayor	0
Regulator	0
Alderman	-0.001
Investor	-0.001
University	-0.002
Client	-0.002
Province	-0.003

Table 5: Random forest MCC scores for individual keywords predicting repayment success, categorized by social capital type (filtered dataset Analysis 2), translated to English

When reviewing the results in table 5, a few keywords stand out. The strongest keywords are found in bonding category, specifically words related to immediate family such as Partner (0.057) and Mother (0.048). Bridging terms showed weak predictive values, with Acquaintance ranking highest with an MCC of 0.009. Linking keywords such as Employer (0.029) and Government official (0.013) demonstrated a moderately positive relation.

5 DISCUSSION

5.1 Discussion

The question this research sought to answer was: 'To what extent can borrowers' social capital -subdivided into bonding, bridging, and linking- predict loan repayment performance in microfinance during the loan screening process?' Contrary to research from Y. Chen et al. (2015) and Iqbal and Rao (2023) which indicated that soft information and

social capital are strong predictors of loan repayment success, this research found a weak, although positive, relation, with even the strongest predictors achieving a performance only slightly exceeding chance level (reflected by the MCC). However, even though the relation appears to be small, due to the size of the dataset and the consistency of the results over multiple configurations, the relevance of social capital as a predictor should not be neglected.

Notably, bonding showed the most potential in the prediction models, contrasting previous literature by Claridge (2018) which indicated that bonding social capital has no effect or even a negative effect on economic outcomes. A possible explanation is that bonding social capital functions differently in the context of Dutch society, which is supported by the fact that Dufhues et al. (2011) found bonding social capital to be a positive predictor of repayment in Thailand but not in Vietnam. This is further embedded in the fundamental subjectivity and context-dependency of soft information, as emphasized by Liberti and Petersen (2018), who also noted that such information is difficult to store and hard to transmit.

Regarding signalling and screening theory, this study assumed that borrowers invested and gave clear cues when social capital was present, to gain an advantage (Connelly et al., 2011). On the other hand, the presumption was made that loan agents would seek and clearly report on this soft information to overcome information asymmetry and increase successful loan outcome (Manove et al., 2001; Zhang et al., 2024). However, the weak predictive strength of social capital could indicate that either the signalling by the borrower is weak or inconsistent, or that the screening done by the loan officers was insufficient or imprecise. This underscores the gap between theory and practice in the way soft information is transmitted and received.

5.2 Practical Implications

Despite the relatively weak overall predictive power of social capital on loan repayment success, this study shows that social capital, especially bonding, can still provide relevant information when predicting repayment success. For example, the stronger predictive keywords could be integrated into an automatic loan decision-making algorithm, improving its accuracy in predicting successful loan repayment. Additionally, standardized procedures could be implemented by financial institutions to support loan agents in detecting and reporting on the presence of social capital, improving its consistency and relevance in loan decision-making and reducing its subjectivity. Finally, this research indicates that the use of soft information -compared to using solely hard information- remains relevant and beneficial in MFI loan decision-making.

5.3 Theoretical Implications

This research contributes to the theoretical understanding of the role that soft information, and in particular social capital, has on repayment success. Using a keyword search to classify soft information or detect social capital have been employed before. However, combining a keyword-based method with the theoretical distinction between the three types of social capital is novel (Aldrich & Meyer, 2015; Nguyen et al., 2013). In addition, this research introduces a method for operationalizing social capital from unstructured textual information using a newly created dictionary, offering a replicable and transparent approach. Furthermore, this research provides insight into the effect that social capital has on loan repayment in a Western country with a developed economy, laying the groundwork for cross-cultural comparisons.

6 LIMITATIONS AND FUTURE RESEARCH

6.1 Limitations

Despite the structured design of this research, several limitations still remain. These limitations include constraints related to the dataset, shortcomings in the method used to extract social capital, and the possibility that social capital might not have a visible significant impact on repayment success.

The dataset's limitations primarily stem from the way loan and risk officers record information related to social capital. Firstly, loan officers might not accurately record details about a borrower's social capital. They might miss or deliberately not note down certain information. Secondly, each officer may document different soft information, and they have varying writing styles. Thirdly, although the overall dataset is sizable, the number of loans that are concluded and are not filtered out is not very large. Furthermore, it is noteworthy how -even in the more filtered dataset- only ~64% of the loans were fully repaid on time, which is lower than expected for a financial institution and could have influenced the study. Finally, the column *Loan repaid (fully or partially) with repayment issues* is ambiguous, as it does not specify the extent of repayment. To consider these loans successful or not, it is essential to know if they are nearly fully repaid, or barely repaid at all.

The next shortcoming of the research is related to the method used to extract the social capital from the unstructured textual data. As seen in the results, the keyword search method used was able to detect the presence of social capital keywords. This, however, does not explicitly mean social capital was present. The method used might have been too lenient in detecting and marking the presence of social capital, thus creating false positives. Furthermore, this method does not take into account the semantic value of the keywords, which could be pivotal in analyzing unstructured text. Finally, even if this method was successful in detecting social capital, it may not have been able to capture the full picture.

Finally, this research did not find a strong relation between social capital and repayment success. This stands in contrast to previous literature, which to a certain degree suggests that social capital does have an impact on financial reliability and loan repayment. One explanation could be that, in the context of micro-credit in the Netherlands, social capital simply plays a less important role in repayment outcome.

6.2 Future Research

Regarding recommendations for future research, a future study could employ more advanced NLP techniques to better capture the context and semantic meaning of soft information, thereby improving the detection of social capital. Moreover, an entirely different approach to identifying the presence of social capital could be used. For example, loan agents could be requested to indicate whether social capital is present following their meetings with clients. Finally, it would be valuable to conduct a similar study in a different country or cultural context to compare the results and assess whether this affects the impact social capital has on loan repayment success.

7 CONCLUSION

The objective of this research was to investigate the predictive power of debtors' social capital, a form of soft information, on loan repayment success. This was done using a keyword search in the program Orange, differentiating between bonding, bridging, and linking based on the network approach. Bonding social capital (strong ties, e.g., close friends and family) was found to have a minor but positive and consistent effect on loan repayment success. The most statistically important individual keywords (Partner, Mother, and (girl)friend) were also found in the bonding category, with 'Employer' from linking social capital ranking fourth. Results from this research contradict some existing literature, primarily by challenging the belief that bonding social capital has little economic relevance and that social capital is a strong predictor overall. The primary limitations of this analysis stem from inconsistencies in the reporting of social capital by loan agents and the use of a rudimentary keyword-based approach applied to unstructured text.

8 ACKNOWLEDGMENTS

Firstly, I would like to express my gratitude to my supervisors Franziska Koefer, MSc, and Dr. Maximilian Goethner, for their support and professional guidance during the writing of this thesis. Furthermore, I would like to thank Dr. Matthias de Visser for his advice on how to conduct the data analysis. Finally, I would like to thank other people close to me for their support and advice during my studies and during the writing of this thesis.

9 REFERENCES

Addae-Korankye, A. (2014). Causes and control of loan default/delinquency in microfinance institutions in ghana. *American International Journal of Contemporary Research*, *4*(12), 36–45. https://www.aijcrnet.com/journals/Vol_4_No_12_December_2014/5.pdf

Aldrich, D. P., & Meyer, M. A. (2015). Social capital and community resilience. *American Behavioral Scientist*, *59*(2), 254–269. https://doi.org/10.1177/0002764214550299

Bafera, J., & Kleinert, S. (2023). Signaling theory in entrepreneurship research: A systematic review and research agenda. *Entrepreneurship Theory and Practice*, *47*(6), 2419–2464. https://doi.org/10.1177/10422587221138489

Baklouti, I., & Bouri, A. (2014). The loan officer's subjective judgment and its role in microfinance institutions. *International Journal of Risk Assessment and Management*, *17*(3), 233–245. https://doi.org/10.1504/IJRAM.2014.062778

Berger, A. N., Miller, N. H., Petersen, M. A., Rajan, R. G., & Stein, J. C. (2005). Does function follow organizational form? evidence from the lending practices of large and small banks. *Journal of Financial Economics*, *76*(2), 237–269. https://doi.org/10.1016/j.jfineco.2004.06.003

Bracht, F., Mahieu, J., & Vanhaverbeke, S. (2024). The signaling value of legal form in entrepreneurial debt financing. *Journal of Business Venturing*, *39*(3), 106380. https://doi.org/10.1016/j.jbusvent.2024.106380

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215. http://www.jstor.org/stable/2676681

Cambria, E., & White, B. (2014). Jumping nlp curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227

Campbell, D., Loumioti, M., & Wittenberg-Moerman, R. (2019). Making sense of soft information: Interpretation bias and loan quality. *Journal of Accounting and Economics*, *68*(2–3), 101240. https://doi.org/10.1016/j.jacceco.2019.101240

Chai, C. P. (2022). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553. https://doi.org/10.1017/S1351324922000213

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

Chen, Y., Huang, R. J., Tsai, J., & Wang, Y. (2015). Soft information and small business lending. *Journal of Financial Services Research*, *47*, 115–133. https://doi.org/10.1007/s10693-013-0187-x

Claridge, T. (2018). Functions of social capital: Bonding, bridging, linking. https://doi.org/10.5281/zenodo.7993853

Connelly, B. E., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, *37*(1), 39–67. https://doi.org/10.1177/0149206310388419

Cornée, S. (2019). The relevance of soft information for predicting small business credit default: Evidence from a social bank. *Journal of Small Business Management*, *57*(3), 699–719. https://doi.org/10.1111/jsbm.12318

Coulthard, B., & Taylor, B. J. (2022). Natural language processing to identify case factors in child protection court proceedings. *Methodological Innovations*, *15*(3), 222–235. https://doi.org/10.1177/20597991221115967

Dube, L., & Verster, T. (2024). Interpretability of the random forest model under class imbalance. *Data Science in*

Finance and Economics, *4*(3), 446–468. https://doi.org/10.3934/DSFE.2024019

Dufhues, T., Buchenrieder, G., Dinh Quoc, H., & Munkung, N. (2011). Social capital and loan repayment performance in southeast asia. *The Journal of Socio-Economics*, *40*(5), 679–691. https://doi.org/10.1016/j.socec.2011.05.007

Gaustad, T., & Bouma, G. (2001). Accurate stemming of dutch for text classification. *Proceedings of the 12th CLIN (Computational Linguistics in the Netherlands)*. https://clinjournal.org/CLIN_proceedings/XII/gaustad.pdf

Herhausen, D., Ludwig, S., Abedin, E., Haque, N. U., & de Jong, D. (2025). From words to insights: Text analysis in business research. *Journal of Business Research*, *198*. https://doi.org/10.1016/j.jbusres.2025.114577

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*, 5979. https://doi.org/10.1038/s41598-022-09954-8

Iqbal, Z., & Rao, Z.-R. (2023). Social capital and loan credit terms: Does it matter in microfinance contract? *Journal of Asian Business and Economic Studies*, *30*(3), 187–209. https://doi.org/10.1108/JABES-10-2021-0185

Jonker, R. A. A., Almeida, T., & Matos, S. (2024). Analyzing a decade of evolution: Trends in natural language processing. In *Lecture notes in computer science* (pp. 162–176, Vol. 14912). https://doi.org/10.1007/978-3-031-68323-7_13

Kim, S., Lessmann, S., Andreeva, G., & Rovatsos, M. (2023). Fair models in credit: Intersectional discrimination and the amplification of inequity. *arXiv preprint arXiv:2308.02680*. https://doi.org/10.48550/arXiv.2308.02680

Kiros, Y. (2020). Loan repayment performance of micro and small enterprises: Evidence from somali region, ethiopia. *Developing Country Studies*, *10*(9), 1–12. https://pdfs.semanticscholar.org/70b2/6d726e86b8050dbd99d1b4933d37454ff08c.pdf

Liberti, J. M., & Petersen, M. A. (2018). Information: Hard and soft. *The Review of Corporate Finance Studies*, *8*(1), 1–41. https://doi.org/10.1093/rcfs/cfy009

Manove, M., Padilla, A. J., & Pagano, M. (2001). Collateral versus project screening: A model of lazy banks. *The RAND Journal of Economics*, *32*(4), 726–744. https://doi.org/10.2307/2696390

McCann, F., & McIndoe-Calder, T. (2015). Firm size, credit scoring accuracy and banks' production of soft information. *Applied Economics*, *47*(33), 3594–3611. https://doi.org/10.1080/00036846.2015.1019034

Muringani, J., Fitjar, R. D., & Rodríguez-Pose, A. (2021). Social capital and economic growth in the regions of europe. *Environment and Planning A: Economy and Space*, *53*(6), 1412–1434. https://doi.org/10.1177/0308518X211000059

Nawai, N., & Shariff, M. N. M. (2010). Determinants of repayment performance in microcredit programs: A review of literature. *International Journal of Business and Social Science*, *1*(2).

Nguyen, T., Dao, B., Phung, D., Venkatesh, S., & Berk, M. (2013). Online social capital: Mood, topical and psycholinguistic analysis. *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM 2013)*, 449–456. https://doi.org/10.1609/icwsm.v7i1.14395

Ntiamoah, E. B., Oteng, E., Opoku, B., & Siaw, A. (2014). Loan default rate and its impact on profitability in financial institutions. *Research Journal of Finance and Accounting*, *5*(14), 67–72.

Paldam, M. (2000). Social capital: One or many? definition and measurement. *Journal of Economic Surveys*, *14*(5), 629–653. https://doi.org/10.1111/1467-6419.00127

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Petersen, M. A. (2004). Information: Hard and soft [CiteSeerX].

Schreiner, M. (2004). *Benefits and pitfalls of statistical credit scoring for microfinance* [Working paper]. https://www.findevgateway.org/paper/2004/12/benefits-and-pitfalls-statistical-credit-scoring-microfinance

Soomro, A., Zakariyah, H., Aftab, S. M. A., Muflehi, M., Shah, A., & Meraj, S. (2024). Loan default prediction using machine learning algorithms: A systematic literature review 2020–2023. *Pakistan Journal of Life and Social Sciences*, *22*(2), 6234–6253. https://doi.org/10.57239/PJLSS-2024-22.2.00469

Wankmüller, S. (2023). A comparison of approaches for imbalanced classification problems in the context of retrieving relevant documents for an analysis. *Journal of Computational Social Science*, *6*(1), 91–163. https://doi.org/10.1007/s42001-022-00191-7

Zhang, J., Shi, W., & Connelly, B. E. (2024). Screening theory and its boundaries: Investigation of screen credibility, necessity, and salience in the context of corporate venture capital. *Academy of Management Journal*, *67*(5), 1359–1391. https://doi.org/10.5465/amj.2021.1185

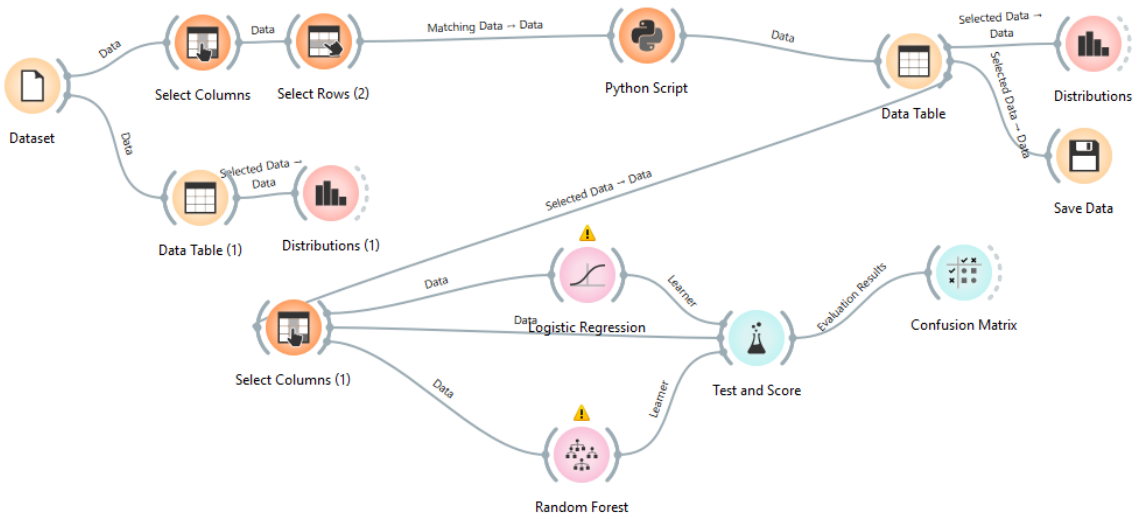
10 APPENDIX

10.1 Social capital keywords table

Bonding	Bridging	Linking
partner	bekende	overheid
vriend	gemeenschap	ambtenaar
vriendin	collega	gemeente
vader	vereniging	provincie
moeder	stichting	school
opa	multicultureel	universiteit
oma	club	rechtbank
schoonvader	sportclub	ministerie
schoonmoeder	buitenland	wethouder
tante	netwerk	burgemeester
oom	competitie	baas
neef	cultuur	leidinggevende
nicht	vrijwilliger	investeerder
zoon	vrijwilligerswerk	sponsor
dochter	intercultureel	client
buren	netwerkbijeenkomst	bestuurder
buurman	buurtcentrum	toezichthouder
buurvrouw		werkgever
huisgenoot		
samen wonen		
kerk		
gezin		
mantelzorger		
echtgenoot		
echtgenote		
kind		
kinderen		
levenspartner		
familie		
vriendengroep		

Table 6: Dutch keywords classified by bonding, bridging, and linking social capital

10.2 Orange Workflow used for predictive modeling



10.3 Python code

```
import csv
import os
import re
import Orange
import numpy as np

# Step 1: Load keywords from Excel-exported CSV
csv_path = r"path/to/keywords.csv"

bonding, bridging, linking = set(), set(), set()
with open(csv_path, newline='', encoding='utf-8') as csvfile:
    reader = csv.reader(csvfile, delimiter=';')
    next(reader)  # skip header
    for row in reader:
        if row[0]: bonding.add(row[0].strip().lower())
        if len(row) > 1 and row[1]: bridging.add(row[1].strip().lower())
        if len(row) > 2 and row[2]: linking.add(row[2].strip().lower())

print("Bonding sample:", list(bonding)[:5])
print("Bridging sample:", list(bridging)[:5])
print("Linking sample:", list(linking)[:5])

# Step 2: Set the column to search
text_field = "Alltext"

# Step 3: Match keywords using word-boundary regex
def get_matches(text, keywords):
    matches = []
    for kw in keywords:
        pattern = rf"\b{re.escape(kw)}\b"
        if re.search(pattern, text):
            matches.append(kw)
    return matches

# Step 4: Process rows
bonding_flags = []
bridging_flags = []
linking_flags = []

for i, row in enumerate(in_data):
    row_text = str(row[text_field]).lower()

    bonding_match = get_matches(row_text, bonding)
    bridging_match = get_matches(row_text, bridging)
    linking_match = get_matches(row_text, linking)

    bonding_flags.append(1 if bonding_match else 0)
    bridging_flags.append(1 if bridging_match else 0)
    linking_flags.append(1 if linking_match else 0)

    if i < 5:  # Debug output
        print(f"\nRow {i} preview:")
        print("Text preview:", row_text[:3000])
        print("Matched bonding:", bonding_match)
        print("Matched bridging:", bridging_match)
        print("Matched linking:", linking_match)

# Step 5: New Orange table with added variables
domain = Orange.data.Domain(
    in_data.domain.attributes + (
        Orange.data.DiscreteVariable("social_capital_bonding", values=["0", "1"]),
        Orange.data.DiscreteVariable("social_capital_bridging", values=["0", "1"]),
        Orange.data.DiscreteVariable("social_capital_linking", values=["0", "1"])
    ),
    in_data.domain.class_vars,
    in_data.domain.metas
)

out_data = Orange.data.Table(
    domain,
    np.hstack((
        in_data.X,
        np.array(bonding_flags).reshape(-1, 1),
        np.array(bridging_flags).reshape(-1, 1),
        np.array(linking_flags).reshape(-1, 1)
    )),
    in_data.Y,
    in_data.metas
)
```