Deciphering Lending Behaviors in Peer-to-Peer Platforms: An Integrated Analysis of Emotion, Topic Modeling, and User-defined Occupational Data

Yunxuan Zhang,

Yunxuanz0306@gmail.com

Bachelor of Arts in Economics, Department of Economics, The University of British Columbia, Vancouver, Canada.

Abstract. This research investigates the influence of borrower-generated content on Peer-to-Peer (P2P) lending outcomes, specifically focusing on the sentiment and thematic content of loan descriptions. The central hypothesis posits that these elements of self-disclosure profoundly influence lending determinations and patterns. Utilizing advanced machine learning techniques, variables such as loan descriptions, occupations, loan amounts, interest rates, and loan statuses are examined. Findings highlight strong correlations between sentiment, thematic quality of loan descriptions, and funding success, accentuating the pivotal role of self-reported occupation. These insights illuminate previously uncharted currents within P2P lending dynamics, emphasizing the importance of sentiment analysis, theme selection, and occupational transparency. The research constitutes a sturdy foundation for prospective investigations in this rapidly evolving field.

Keywords: Peer-to-Peer (P2P) Lending, Self-Disclosure, Sentiment Analysis

1 Introduction

In recent years, the advent of Peer-to-Peer (P2P) lending platforms has instigated a paradigmatic transformation in the financial industry by facilitating direct transactions between lenders and borrowers. This effectively circumvents the standard banking systems thereby democratizing the lending environment. As Bachmann et al. (2011) [1] elucidated, borrowers could access loans more advantageously under this model, often securing more favorable conditions. Drawing insights from Mushtaq and Bruneau's research (2019) [2], such digitally mediated financial practices could potentially impact poverty and inequality reduction, thereby necessitating comprehensive examination in this sphere. The unconventional and personalized nature of P2P loans, usually characterized by borrower-provided loan descriptions, unfolds unique opportunities and challenges for risk assessment.

Loan descriptions and user-defined occupations, significant yet under-explored aspects, offer potential insights into creditworthiness and repayment capabilities. The complexity of natural language data analysis may explain the existing literature gap. Utilizing Natural Language Processing (NLP) and topic modeling, the predictive value of these unstructured data for successful loan transactions can be uncovered.

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345291

2 Literature review

Since the inception of the initial P2P lending platform, Zopa, in 2005, an abundant corpus of academic literature has surfaced. Much research pivots around the leading U.S. P2P lending platform, Prosper, which has made its data available to the public.

Traditional lending decisions largely hinged on institutional risk assessments, primarily led by banking institutions, with credit scores and assets acting as prime determinants. However, the advent of P2P lending has prompted a transition towards a more personalized lending model. In this context, information furnished by the borrower plays an integral role in gauging creditworthiness. This encompasses a variety of elements such as loan descriptions, self-reported occupations, endorsements from acquaintances, and the nature of the borrower-lender relationship (Lin, Prabhala, & Viswanathan, 2013) [3]. This holistic perspective affords lenders a more comprehensive understanding of the borrower's financial behavior and credibility, an aspect typically overlooked in traditional lending frameworks.

Highlighting an empirical relationship between the narrative quality of loan applications—which incorporated elements such as readability, positivity, and deceptive cues—and default probabilities, Gao, Lin, and Sias (2022) [4] underscored the importance of soft information in P2P lending practices. Further, user-generated short texts significantly bolstered the prediction of credit defaults, thus underscoring the importance of borrowers' self-reported loan purposes and textual features in creditworthiness assessment (Kriebel & Stitz, 2022) [5]. These insights collectively highlight the nuanced and crucial role of soft information in P2P lending, advocating for a multi-dimensional approach to creditworthiness evaluation that extends beyond conventional financial metrics.

3 Methodology

3.1 Data Collection

The dataset employed for this study was derived from Prosper.com, a leading U.S. P2P lending platform that avails its data for public use. This dataset comprises information collected from 2005 to 2008, featuring a broad spectrum of loan applications, lending determinations, and repayment statuses. It consists of a sum of 582,982 entries, with each entry corresponding to a distinct loan application, and it encompasses 20 variables that provide detailed insights into divergent facets of the lending process. Notably, the original data consists of a balanced mix of 10 numerical and 10 categorical variables respectively.

3.2 Feature engineering and text processing analysis

The dataset was subjected to an extensive preprocessing routine to ready it for subsequent analytical procedures. This included handling missing data, managing outliers, and standardizing numerical features, as required. Emphasis was given to the 'loan_description' variable, processed using the Natural Language Toolkit (NLTK) as suggested by Bird et al. (2009) [6]. By employing these sophisticated natural language processing techniques, a sentiment score was crafted which encapsulated the emotional tone of loan applications, paving the way for more nuanced insights into the borrower's narrative.

To decipher the concealed semantic structures inherent in the loan descriptions, an adaptive model selection approach for Latent Dirichlet Allocation (LDA) was adopted, drawing on the density-based method proposed by Cao et al. (2009) [7]. As an advanced unsupervised machine learning technique, LDA enables the extraction of implicit topics from a large corpus of text. Applying this approach, a total of 14 distinct topics were identified from the loan descriptions, each encapsulating a unique aspect of the borrowing narrative. Adhering to the guidance from Cao et al. (2009) [7], each loan description was subsequently correlated with topic probabilities, conveying the extent of their affiliation with the derived topics. The entire process is illustrated in Figure 1.

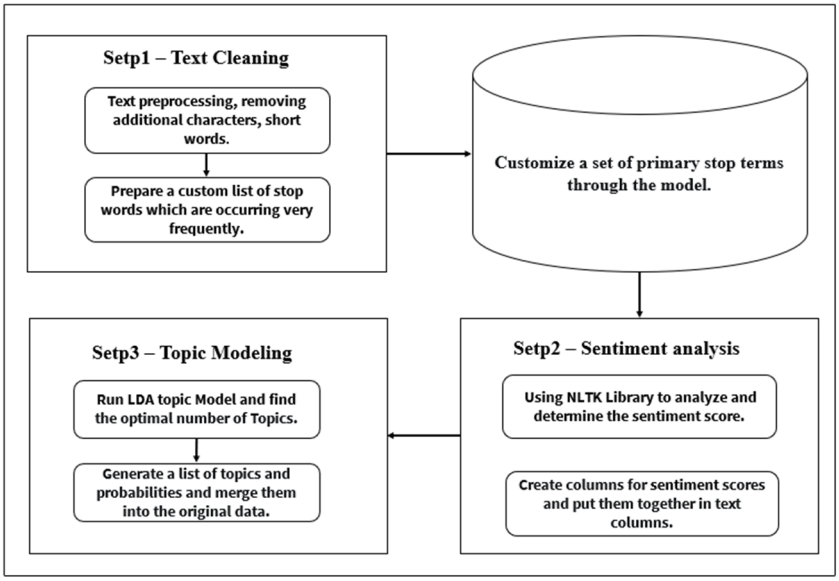

Fig. 1. Processing Flow for 'Loan_Description' Text Column

3.3 Regression analysis

In the ensuing stage of the research, regression analysis was utilized to scrutinize the relationships between the independent variables—comprising sentiment scores, topic probabilities, time, and occupation—and the dependent variable, the loan funding success. The dataset was augmented with topic probabilities, revealing underlying themes embedded in the loan descriptions. Each description was associated with its most probable topic and the corresponding probability score.

$$\log(Bidamount) = C + \beta 1 \cdot Sentiment\_\text{Score} + \beta 2 \cdot Topic\_Probility + \beta 3 \cdot Occupation + \cdots + \beta 4 \cdot Other\ control\ variables + \varepsilon i \quad (1)$$

Upon completion of dataset assembly, a regression model was constructed and trained, permitting an extensive examination of the relationships between the variables and the success of the loan funding. This in-depth regression analysis could offer valuable insights into the influence

of these features on funding success, thereby playing an integral role in deciphering the dynamics of P2P lending platforms.

4 Result

4.1 Dependent and independent variables

The dependent variables, independent variables, and control variables used in the analysis are shown in Table 1.

Table1. Main variables and descriptions.

Variable	Description
Dependent variable	
Bidamount	The amount received by borrower.
Independent variable	
Sentiment score	Score of sentiment polarity.
Occupation with One-hot-encode	Employment categories.
Dummy_encoding time	Temporal instances transformed.
Topic_probabilities	Probabilities associated with topics.
Other control variables	Further details provided later.

4.2 Regression result

The following section presents a thorough interpretation of the regression results. The regression aimed to distinguish the statistically significant predictors from the array of independent and control variables for the dependent variable - the bid amount These findings elucidate the influence of each variable on investment behavior and yield valuable insights into the operations of peer-to-peer lending. For a more detailed overview of these results, reference can be made to Table 2&3.

Table 2. OLS Regression Results (Constant)

Variable	Coef.
Constant	72.1086**

*Note: statistical significance, with ** indicating $p < 0.05$, and * indicating $p < 0.10$.*

Table 3. OLS Regression Results (Sentiment, Topics)

Variable	Coef.	Variable	Coef.
Sentiment score	1.2490*	Investment	18.1127*
Expenses	15.3082	Home	25.6188**
Assistance	19.2612*	Event	36.7429
Loans	22.8394**	Property	12.1016
Credit	20.1431*	Education	21.7642**
Monthly expense	19.9690*	Debt	32.3269**

*Note: statistical significance, with ** indicating $p < 0.05$, and * indicating $p < 0.10$.*

Table 3 underscores the statistically significant variables derived from the regression analysis. By scrutinizing these temporal and occupational factors, a richer understanding of their roles and implications on the dependent variable, the bid amount, is attained. This approach hence provides an opportunity to gain deeper insights into the complexities inherent in peer-to-peer lending, allowing for more nuanced conclusions to be drawn.

Table 4. OLS Regression Results continue (Time and Occupation Analysis)

Variable	Coef.	Variable	Coef.
month_2	9.5894**	month_3	6.0877**
month_4	4.3502**	month_5	3.8078**
month_6	-3.2298**	month_7	10.6075**
month_8	8.6121**	month_10	7.5179**
momth_12	-14.0309**	Accountant/CPA	3.1764*
Analyst	3.0472*	Clerical	2.8708*
Not Applicable	31.3947**	Other	3.8841**
Professional	2.6289**	Retail	3.1262*

*Note: statistical significance, with ** indicating $p < 0.05$, and * indicating $p < 0.10$.*

5 Discussion and Conclusion

The subsequent discussion interprets regression analysis results in the context of peer-to-peer lending. Our analysis highlights a positive correlation between sentiment scores and bid amounts (Loughran & McDonald, 2010) [8], hinting at the influence of emotional resonance in loan descriptions on investor behavior.

Remarkably, empirical evidence indicates that when P2P lending platforms refrain from disclosing the occupations of individual borrowers, it might result in a surge in lending activities. This challenges the long-standing conventional emphasis on transparency in the financial sector (Hermalin & Weisbach, 2012) [9]. This observation implies that withholding certain information could possibly mitigate biases and generate higher bids.

Regarding topic selection, education and credit cards emerge as favorable choices, as they have the potential to enhance borrowers' perceived credibility and intentions. Additionally, the lending industry exhibits seasonality patterns, with reduced activity observed at the year-end and peak activity occurring in July, mirror investment behaviors documented by Dichev and Janes (2003) [10]. These findings underscore the potential of sentiment analysis, temporal patterns, and information disclosure in improving the efficiency of borrower-lender matching algorithms and shaping the evolving landscape of peer-to-peer lending.

In conclusion, this study elucidates decision-making intricacies within the peer-to-peer lending ecosystem, demonstrating how sentiment scores, professional disclosure, choice of loan description topics, and temporal factors subtly sway lending decisions. It is observed that within this P2P lending microcosm, not only do quantitative elements matter, but the qualitative aspects—the narratives and emotions—also significantly influence both lenders and borrowers. These insights provide valuable strategic guidance for borrowers and platforms alike, underscoring P2P lending's role as a compelling alternative to traditional financial avenues.

References

[1] Bachmann, A., Becker, A., Buerckner, D., Hilker, M., Kock, F., Lehmann, M., Tiburtius, P., & Funk, B. (2011). Online Peer-to-Peer Lending--A Literature. Journal of Internet Banking and Commerce, 16.

[2] Mushtaq, R., & Bruneau, C. (2019). Microfinance, financial inclusion and ICT: Implications for poverty and inequality. Technology in Society, 59, 101154. https://doi.org/10.1016/j.techsoc.2019.101154.

[3] Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending. Management Science, 59(1), 17–35. http://www.jstor.org/stable/23359603.

[4] Gao, Q., Lin, M., & Sias, R. (2023). Words Matter: The Role of Readability, Tone, and Deception Cues in Online Credit Markets. Journal of Financial and Quantitative Analysis, 58(1), 1-28. doi:10.1017/S0022109022000850.

[5] Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. European Journal of Operational Research, 302(1), 309-323.

[6] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

[7] Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing, 72(7–9), 1775-1781.

[8] Loughran, T., & McDonald, B. (2010). When is a liability not a liability? Textual Analysis, Dictionaries, and 10-Ks. Journal of Finance, Forthcoming. Available at SSRN: https://ssrn.com/abstract=1331573.

[9] Hermalin, B. E., & Weisbach, M. S. (2012). Information disclosure and corporate governance. The journal of finance, 67(1), 195-233.

[10] Dichev, I. D., & Janes, T. D. (2003). Lunar Cycle Effects in Stock Returns. The Journal of Private Equity, 6(4), 8–29. http://www.jstor.org/stable/43503349.