Prediction of Default Risk in Peer-to-Peer Lending Using Structured and Unstructured Data

Jei Young Lee[a]

Using data from Lending Club*, we analyzed funded loans between 2012 and 2013, the default status of which were mostly known in 2018. Our results showed that both the borrower characteristics and the conditions of the loan were significantly associated with the loan default rate. Results also showed that the sentiment of a user-written loan description influenced the borrower's loan interest rates. It contributes to expanding the scope of peer-to-peer (P2P) loan research by implementing unstructured data as a new model variable. Financial counselors need to consider the growth potential of the P2P loan market using data analysis: This will reveal niche market opportunities, enabling the development of services necessary for the safe supply of small loans at reasonable interest rates.*

Keywords: endogeneity, instrumental variable method, P2P loan default risk, probit model, sentiment analysis

Peer-to-peer (P2P) lending allows individuals to lend money to other individuals conveniently via online platforms. Without the intervention of a financial institution, a P2P online marketplace connects borrowers and lenders directly. As a new e-commerce phenomenon in the financial field, P2P lending is considered a potential method of increasing economic efficiency (Berger & Gleisner, 2009; Guo, Zhou, Luo, Liu, & Xiong, 2016; Lin, Li, & Zheng, 2017). The costs saved from reduced intervention from financial institutions can result in a higher return for lenders and lower interest rates for borrowers (Serrano-Cinca, Gutiérrez-Nieto, & López-Palacios, 2015). Low operational costs, simplified and quick procedures are other big advantages of P2P lending platforms. Due to this competitive advantage over conventional banks, the P2P lending market is expected to grow rapidly (Garret, 2017). It is estimated that the global P2P lending industry will be worth $290 billion by 2020 (Heber, 2015).

While P2P lending platforms offer many options for lenders, it is not easy for an individual lender to choose the right borrowers to invest in. Information asymmetry between lenders and borrowers exists (Serrano-Cinca & Gutiérrez-Nieto, 2016), and lenders can be uninformed or have inaccurate information about borrowers and their loan characteristics. This information asymmetry problem can lead to adverse selection, and an individual lender may erroneously choose borrowers with a high probability of default. In a pseudonymous online environment, the information asymmetry between lenders and borrowers becomes even more acute (Pokorná & Sponer, 2016). Certainly, loan default risk would be reduced as more information on borrowers is obtained. This information may include not only the creditworthiness of the borrowers, but also their current financial status and requested loan characteristics. Based on a variety of criteria, an individual lender needs to consider the risk of loan default to make an investment decision.

To help lenders and investors make optimal decisions, P2P lending companies provide information on the credit history and financial status of the borrower (e.g., annual income, home ownership, debt-to-income [DTI] ratio, and revolving credit balance) and corresponding loan characteristics (e.g., loan amount, loan term, interest rates, and a description of the loan purpose). Using this information to predict loan default risk correctly is an important concern for P2P lenders because they bear much higher risk than financial institutions, which specialize in risk management (Serrano-Cinca et al., 2015). In addition, small loan opportunities are increasing as the P2P lending market substitutes for and

[a]*Assistant Professor, Department of Business Administration, Gwangju University, Gwangju 61743, Republic of Korea. E-mail: jylee@gwangju.ac.kr*

http://dx.doi.org/10.1891/JFCP-18-00073

complements the conventional bank-centered lending service market. However, the growing number of low-quality borrowers migrating to the P2P loan market is simultaneously increasing the risk and management difficulties associated with such loans (Tang, 2019). The possibility of misinterpreting a borrower's creditworthiness leads to lenders' lower profits. Moreover, managing the loan default risk is also critical for P2P lending companies to invigorate the company's own platform. This necessitates the development of a more accurate predictive model of default risk in the P2P lending and micro-loan market.

In this study, therefore, we attempted to further develop a loan risk model by analyzing both structured and unstructured data. Using data from a large U.S. P2P lending company, *Lending Club*, we empirically explored the factors associated with loan default risk and investigated whether the sentiment of unstructured data (i.e., a user-written text of the loan purpose) improved our predictive model. Because we assumed that unobserved shocks could affect both the P2P loan interest rate and its default risk, an instrumental variable (IV) approach was used to rule out an endogeneity problem in our predictive model. We found that both the borrower characteristics and the conditions of the loan were significantly associated with the loan default rate. Results also showed that the sentiment of a user-written loan description influences the borrower's loan interest rates. P2P loan interest rates was one of the critical loan conditions for borrowers and directly affected loan default risk. Our study suggests that customized financial advice provided to borrowers and investors should be based on data analysis. The loan default risk model suggested in this study enables financial counselors to quantitatively explain how much effect each model variable has on the P2P loan interest rates and the default probability. In addition to proposing a new model (i.e., an IV probit model) that reduces the bias in the existing loan default prediction estimation model, the current study contributes to the scope expansion of P2P loan research by implementing text-based unstructured data (i.e., loan description as a new variable).

Related Literature and Hypotheses

Recent studies on P2P lending have investigated how P2P lending markets operate and the factors that affect borrower default risk and lender profitability. Berger and Gleisner (2009) conducted one of the first empirical studies on P2P lending and analyzed the role of financial intermediaries using data from a P2P lending platform, Prosper.com. Emekter, Tu, Jirasakuldech, and Lu (2014) expanded Berger and Gleisner's study and developed a model to evaluate loan credit risk and performance. They found that the borrower's credit grade, DTI ratio, FICO (Fair Isaac Corporation) score, and revolving line utilization impacted loan default. Guo et al. (2016) also focused on developing a credit risk assessment model, which evaluated the return and risk of each individual loan. The loan attributes used in their model included the amount of the loan, the borrower's DTI and home ownership status. Similarly, Lin et al. (2017) also explored determining factors of default risk based on the demographic characteristics of borrowers and corresponding loan information. Another study by Serrano-Cinca et al. (2015) analyzed the relationships among the borrower's grade (assigned by the P2P lending company), the loan interest rate, and the probability of default. Their results also suggested that loan characteristics and borrower indebtedness were related to default risk. Interestingly, they found that the higher the interest rate, the higher the probability of default. The current study pays attention to interest rates, which may indicate the probability that such P2P loans will default and are simultaneously connected to the profit rates of P2P lending platforms and investors. In addition, structured and unstructured variables that can affect these elements will be examined. Among the many explanatory variables that influence the loan default risk, the endogeneity problem harbored by loan interest rates will be diagnosed as well. In sum, our study is differentiated from its predecessors in that a binary probit model with the IV method is used to resolve the estimation bias problem of existing loan default risk prediction models caused by endogenous regressors, in addition to investigating the variables that affect the loan interest rates determined by P2P lending companies.

> **H1:** P2P loan characteristics and borrower characteristics are associated with both loan interest rates and loan default risk.
>
> **H2:** The distribution of P2P loan interest rates is the same between default and non-default loans.

As the P2P lending market grows, the scope of research about it is also expanding. A more recent study by Han, Xiao, and Su (2019) focused on the financial knowledge

and risk attitude of the borrower as factors that can influence P2P borrowing behaviors. In addition, it demonstrated that borrowers who exhibit familiarity and expertise about finance, as well as a risk-taking tendency, are more likely to apply for a P2P loan. Chen, Jiang, and Liu (2018) showed that the loan performance (i.e., loan default rate and loan return) could vary depending on gender difference of the investor. Furthermore, the study showed that gender difference can be mitigated by the income, education level, and occupation of the investor. Previous studies, however, are limited in that they focused on structured data (i.e., numbers or dummies) such as borrower or loan characteristics (that are measured via structured methods), and predicted loan default risk without considering unstructured data (i.e., a user-written text or description).

The information available to P2P lenders also includes a description of the requested loan written by the borrower. Some major P2P lending companies, such as *Lending Club*, evaluate each loan by grade, loan purpose, and loan description (Nowak, Ross, & Yencha, 2018). To utilize the full value of our given data, we extracted the loan description and analyzed it with sentiment analysis. In contrast to structured data, unstructured data, such as texts, cannot be used directly for model estimation without preprocessing. Sentiment analysis focuses on extracting emotions (i.e., positive, neutral, and negative) from online text and representing these extractions with numeric values (Liu, 2012). Thus, researchers can extract a user's emotional attitude on a specific topic of interest with sentiment analysis, and this approach is especially valuable to researchers given the growing use of user-generated contents (UGC) on online platforms (Rambocas & Pacheco, 2018).

Sentiment analysis extracts the emotional expression language used in the text and measures the degree of polarity present in the written text as positive, negative, or neutral (e.g., -2: very negative, -1: somewhat negative, 0: neutral, 1: somewhat positive, 2: very positive). In this process, the collected text data are compared with existing emotional lexicons and assigned numeric values, which quantify the sentiment score (Kang & Park, 2014). Thus, the emotional lexicons serve as a criterion for providing a polarity category and sentiment score for each emotive phrase. In our study, the total sentiment score for the loan purpose description written by a user was calculated by multiplying each sentiment score with its weight (i.e., if there were adverbs in the sentence, such as "too," "enough," "very," and "extremely") and then summing them up. In addition to structured data, these unstructured data in words and sentences can also provide new information for P2P loan companies to make their loan decisions.

Several studies about the financial market have used sentiment analysis to predict stock market performance. Tirunillai and Tellis (2012) examined whether a relationship existed between UGC and the stock market performance of the firm. Using consumer reviews from online shopping websites, these authors found that negative and positive UGC had asymmetric effects on stock returns. Another study by Chen, De, Hu, and Hwang (2014) also extracted investor opinions from the social media website *Seeking Alpha* and predicted future stock returns and earnings surprises. These results showed that the fraction of negative words in articles and comments were strongly related to earnings surprises. Finally, Schweidel and Moe (2014) modeled the sentiment expressed in social media posts across venues and investigated whether there existed significant variations in sentiments. The findings in that study also suggested that a measure of brand sentiment can serve as a leading indicator of change in the firm's stock price. All these studies have used sentiment analysis as a tool for social media and financial market research.

Thus, our study attempted to predict loan default risk by using sentiment analysis as a new evaluation tool, together with existing structured data. To extend the current research on P2P lending, we tried to find a causal relationship between explanatory variables (i.e., borrower characteristics and the conditions of a loan, such as interest rates, loan amount and term, loan purpose) and a response variable (i.e., loan default) by using both structured and unstructured data. To analyze the text variable, loan description, we used sentiment analysis to classify the sentences as conveying a positive, negative, or neutral emotion. This polarity classification was expected to create the borrower's sentiment variable and enhance our loan default prediction model.

Meanwhile, the loan interest rate is one of the critical factors affecting a lender's investment decision, and some P2P lending companies (e.g., *Lending Club, Prosper*) set each loan's interest rate based on the borrower's overall grade. Theoretically, the interest rate at which lenders are willing to invest is influenced by the level of default risk (Mild, Waitz,

& Wöckl, 2015). This implies that the borrower's creditworthiness and loan-specific characteristics, which affect the probability of default, can also influence the loan interest rate. In applied econometrics, this can cause an endogeneity problem. If any regressor is endogenous within a model, then the estimates of all regression parameters will be biased and inconsistent, and this endogeneity problem is quite likely to occur when cross-sectional or observational data are used (Cameron & Trivedi, 2005). We began our analysis by assuming that unobserved shocks affect both the P2P loan interest rate and its default risk, and thus, there exists an endogenous effect of the loan interest rate.

> **H3:** The sentiment of a P2P loan description is associated with both loan interest rates and loan default risk.
>
> **H4:** The distribution of P2P loan interest rates is the same between positive and negative sentiment groups.

Our study aimed to empirically investigate the factors influencing the default risk of P2P loans and tried to accurately predict loan default risk. As the goal was to present a predictive model for addressing endogeneity between factors, we used a binary probit model with an IV approach. To the best of our knowledge, this study is the first to consider both users' sentiment and endogenous effects in P2P lending credit risk modeling. We expected our results to suggest a novel approach to predict loan default risk more accurately and provide significant insights into which loan-related factors should be considered when assessing online P2P loans.

Data

Our study utilized data from *Lending Club*, one of the world's largest P2P lending platform in loan issuance. The company's lending site provides historical loan data including borrower's financial situation, loan characteristics, and current loan status (i.e., current, late, paid in full, default). Because the maximum maturity of the company's loan is 60 months, we analyzed all funded loans between 2012 and 2013 so that the status of most loans (i.e., default or non-default) were known in 2018. Following recent research on P2P loan default, we assumed that non-defaulted loans were "paid in full" or on the payment schedule (see Lin et al., 2017), whereas defaulted loans were late in payment or officially depreciated. The final sample in our analysis contained 188,181 loans, of which 28,933 were default.

As there could be multicollinearity between explanatory variables, we selected the most commonly used variables in credit risk assessment models for our analysis. The variables used for our proposed model were broadly categorized into two groups: *loan characteristics* and *borrower characteristics*. For loan characteristics, we used loan interest rates, loan amount (10,000s of dollars), loan term (36 months or 60 months), and dummies for loan purpose (e.g., car, credit card, debt consolidation; see Table 3). For sentiment analysis, we also extracted a sentiment score and counted the total number of words of each loan description. For borrower characteristics, we used the borrower's annual income (10,000s of dollars), a dummy for home ownership, DTI ratio (%), revolving credit balance (10,000s of dollars), number of past-due delinquencies, and borrower's overall grade assigned by the P2P lending company. Among the variables, the sentiment score and the total number of words of loan description, the borrower's overall grade, and the number of previous delinquencies were used as IVs in our predictive model. In our context, these four variables were considered to determine the loan interest rate but not to affect loan default risk directly and could thus be regarded as appropriate instruments. IV methods that use lagged values or calculate a proxy variable have been common in econometric analysis (see Chintagunta, Gopinath, & Venkataraman, 2010; Hendricks, Janzen, & Smith, 2015; Petrin & Train, 2010). In consumer finance studies, Shin and Kim (2018) also examined the relationship between the subjective income risks and stock ownership on behalf of households in the United States during the Great Recession and used unemployment rates by industry as an IV in their study. Table 1 shows summary statistics for the variables used in our study.

Because we were also interested in whether both loan interest rates and loan default risk could be dependent on the sentiment of a loan description, we conducted a *t* test for equality of means between positive and negative sentiment groups. The sentiment score of the loan description was computed by summing all positive and negative values of each sentence in the description. For the *t* test, we assigned a loan into the positive group if the sentiment score was greater than 0, or otherwise assigned a loan into the negative group if the sentiment score was less than 0. Table 2 shows the results of the *t* test for equality of means between different sentiment groups. It showed a significant difference in the means of positive and negative sentiment groups.

TABLE 1. Summary Statistics

Variable	Mean	Std. Dev.	Min	Max
Loan default (1 = default, 0 = non-default)	0.154	0.361	0	1.00
Interest rate (%)	14.278	4.437	6	26.06
Loan amount (10,000s of dollars)	1.435	0.811	0.1	3.5
Loan term[a]	0.235	0.424	0	1.00
Annual income (10,000s of dollars)	7.223	5.182	0.48	714.18
Home ownership[b]	0.082	0.275	0	1.00
DTI ratio (%)	17.060	7.598	0	34.99
Revolving balance (10,000s of dollars)	1.632	1.929	0	256.90
Sentiment score of loan description	0.098	0.307	–3.436	5.105
Total words of loan description	12.702	21.670	0	1.69
Borrower's overall grade[c]	5.243	1.288	1	7.00
Number of borrower's previous delinquencies	0.240	0.704	0	29.00

Note. DTI = debt-to-income.
[a]Loan term of 36 months and 60 months is assigned 0 and 1, respectively. [b]Home ownership as a dummy, where 1 = own, 0 = rent, mortgage or other. [c]Grade is divided into A, B, C, D, E, F, G, and they are assigned value 7, 6, 5, 4, 3, 2, 1, respectively. Summary statistics for dummies (i.e., loan purposes) are omitted for simplicity.

TABLE 2. Results of *t* Test for Equality of Means Between Different Sentiment Groups

	Positive		Negative		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	*t*-Test (Negative-Positive)
Loan default	0.148	0.355	0.164	0.370	0.016***
Interest Rate (%)	13.607	4.392	14.343	4.484	0.736***

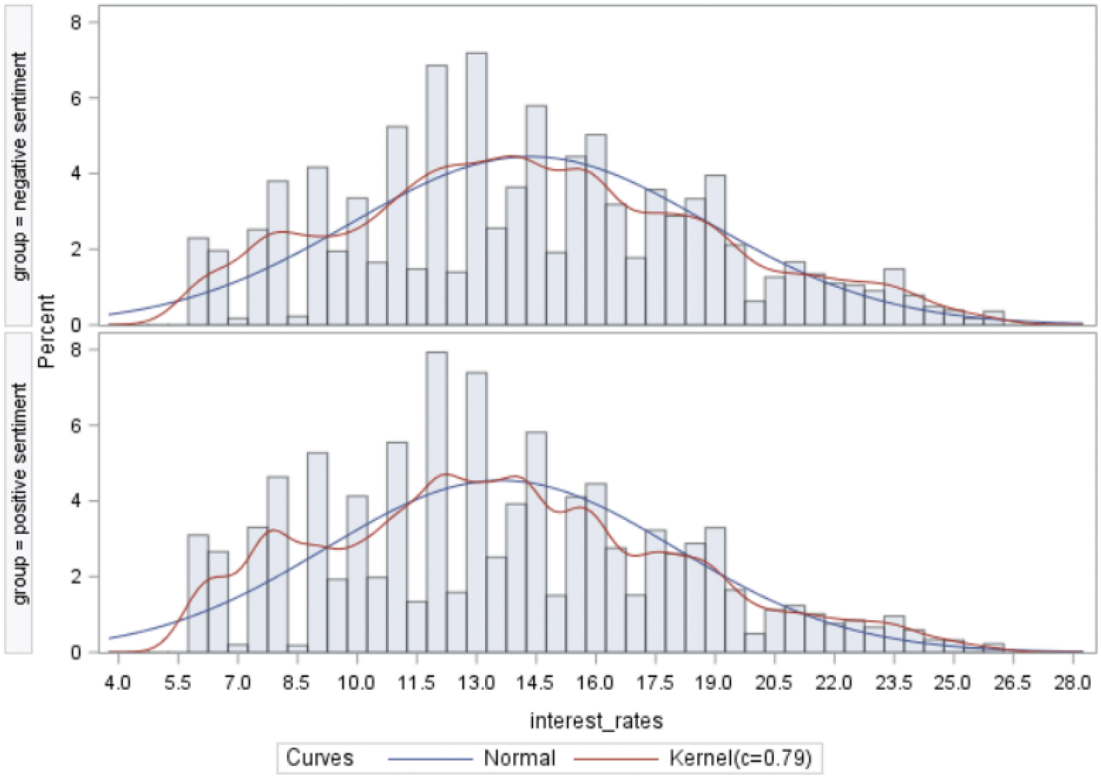
***$p < .01$.

We found that a loan with positive sentiment tended to have a lower interest rate and a lower level of default risk. This implies that a borrower's overall sentiment expressed by a loan description can be highly related to loan conditions (i.e., interest rates) and the probability of loan repayments (i.e., loan default risk). Results of the *t*test support our hypothesis that the sentiment of a P2P loan description is associated with both loan interest and loan default risk (i.e., it supports hypothesis 3).

To have a better sense of how the sentiment score of a loan description, loan interest rates, and loan default risk are associated with each other, we plotted the distribution of loan interest rates depending on (a) sentiment score (i.e., positive or negative) and (b) loan status (i.e., defaulted or non-defaulted). Figures (1 and 2) show that loan interest rates have a strong association with both sentiment score and loan status; that is, loans with negative sentiment generally have higher interest rates (Figure 1), and loans with higher interest rates are more likely to default (Figure 2).

As shown in Figure 1, the histogram of the positive sentiment group is more right-skewed than that of the negative sentiment group, which means that the positive sentiment group generally has lower interest rates. Similarly, Figure 2 shows the interest rate distribution by loan default status. The histogram of defaulted loans is more left-skewed, which means defaulted loans generally have higher interest rates. This finding is consistent with a previous study that showed that an increase in interest rates usually worsens a borrower's financial situation and can result in higher future defaults (Beutler, Bichsel, Bruhin, & Danton, 2017). The Kolmogorov–Smirnov test was also conducted to see if there was a significant difference in the distribution of interest rates by sentiment score and loan default status. Our test rejected the null-hypothesis that the distribution of interest rates is the same between the two groups for both the sentiment score and loan default status variables (the *KS* statistic = 0.0323, p-value < .0001 for sentiment score; and the *KS* statistic = 0.0841, p-value < .0001 for loan default status). This finding indicates a statistically significant difference in

Figure 1. Histogram of interest rates by sentiment score.



Note. The *p*-value for the Kolmogorov–Smirnov two-sample test is < .0001 (the *KS* Statistic = 0.0323), which indicates the interest rates across sentiment groups do not have the same statistical distribution.

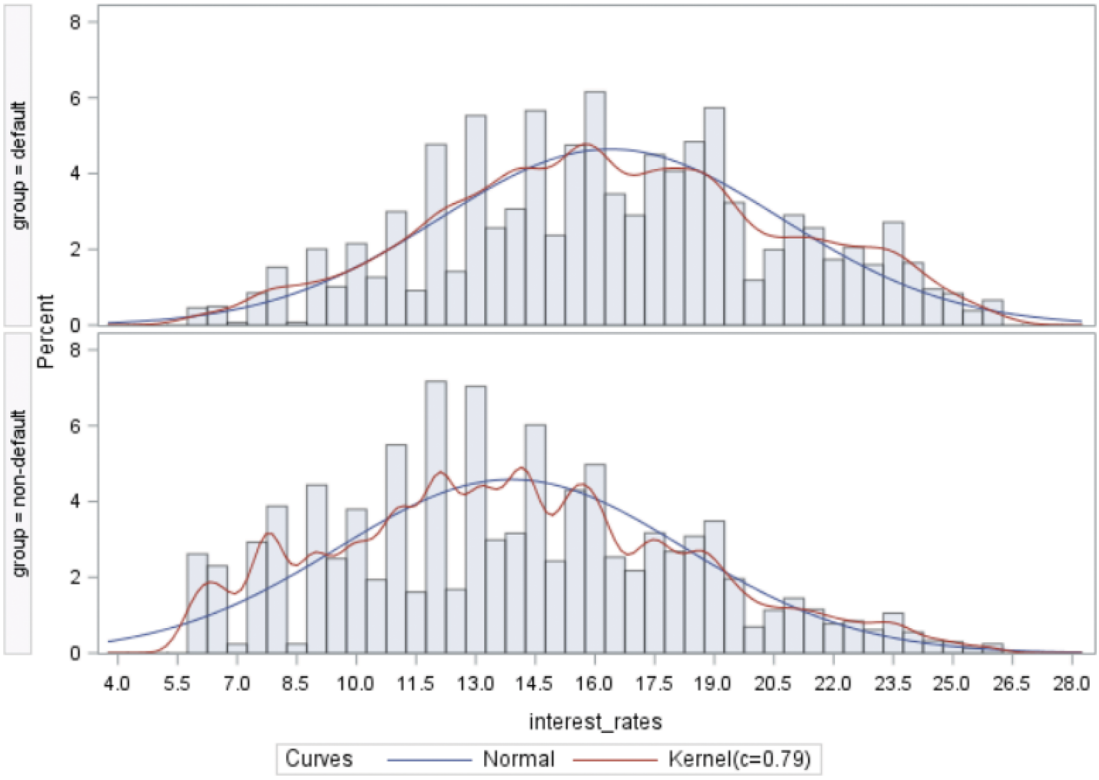
the distribution of the interest rates between the two groups for each variable. Results of the Kolmogorov–Smirnov test, therefore, reject our hypotheses that the distribution of P2P loan interest rates is the same (a) between default and non-default loans, and (b) between positive and negative sentiment groups (i.e., it rejects both hypotheses 2 and 4).

Model

To measure the exact effect of each variable on loan default risk, we used a binary probit model with the IV method. The exclusion of an explanatory variable that can affect a dependent variable (i.e., omitted variable bias) is the most common cause of the endogeneity problem and is also difficult to diagnose in the model (Rutz & Watson, 2019). Omitting variables that can affect the dependent variable results in an endogenous problem due to the correlation between the error term and explanatory variables in the model. This omitted variable problem can be caused by unavailable data in the model building process. In our context, this omitted variable bias can cause the endogeneity issue of our "interest rates" variable. P2P loan companies determine the loan interest rate by taking into account both the borrower's financial status and the requested loan condition, as well as the situation of the entire loan market (e.g., the overall economic condition, consumer demand for P2P loans, degree of market competition). For example, Dietrich and Wernli (2016) suggested that in addition to loan-specific and borrower-specific factors, other macroeconomic factors and the number of loan-auctions conducted before the loan is granted can also drive the interest rates in the P2P consumer lending market. However, in general, a researcher cannot completely observe a firm's own mechanism of determining loan interest rates, and, thus, the interest rates that a researcher identifies through an empirical dataset are non-random.

In many econometric studies, an IV approach is used as a way to address the omitted variable bias. For the model to be correctly estimated through the IV method,

Figure 2. Histogram of interest rates by loan status.



Note. The *p*-value for the Kolmogorov–Smirnov two-sample test is < .0001 (the *KS* Statistic = 0.0841), which indicates the interest rates across loan status groups do not have the same statistical distribution.

it should satisfy both the conditions of "instrument relevance" and "instrument exogeneity" (Angrist, Imbens, & Rubin, 1996; Rutz & Watson, 2019). That is, our IVs should be independent of the loan default rate and have a high correlation with interest rates. In this respect, we selected four IVs: the borrower's overall grade, the borrower's number of delinquencies over the past 2 years, the sentiment score, and the total number of words of the requested loan description. As the grade and number of past delinquencies are measured based on the credit history of the loan applicant to date, these variables can affect the level of loan interest rates but are weakly related to the default rate of the newly applied P2P loan. In addition, the sentiment score and total number of words extracted from the unstructured data of the loan purpose description can be considered as qualitative information when P2P firms are determining the loan interest rate; however, it is reasonable to assume the loan purpose description itself does not directly affect loan default risk.

Because we assumed that the loan interest rate is endogenous and correlated with the error term, we modeled the loan interest rate with four IVs—the sentiment score and the total number of words of loan description, the borrower's overall grade and the number of delinquencies over the past 2 years. Formally, we consider the following model in which our dependent variable (y) is binary (i.e., default or non-default).

$$y_i^* = \beta_0 + \beta_1 IntRate_i + \vec{\beta}_2 \cdot \vec{LoanCH}_i + \vec{\beta}_3 \cdot \vec{BorrowerCH}_i + \vec{\beta}_4 \cdot \vec{Controls}_i + u_i \tag{1}$$

$$\begin{aligned} IntRate_i = {} & \gamma_0 + \gamma_1 Sentiment_i + \gamma_2 TotalWords_i + \gamma_3 Grade_i \\ & + \gamma_4 PastDelinq_i \\ & + \vec{\gamma}_5 \cdot \vec{LoanCH}_i + \vec{\gamma}_6 \cdot \vec{BorrowerCH}_i \\ & + \vec{\gamma}_7 \cdot \vec{Controls}_i + v_i \end{aligned} \tag{2}$$

where *i* indexes the loan requested by each borrower. The dependent variable y_i^* is latent. We observe a binary outcome y_i, with $y_i = 1$ if $y_i^* > 0$ (if a loan *i* default), and $y_i = 0$ if $y_i^* \leq 0$ (if a loan *i* non-default). $IntRate_i$ represents the interest rate of loan *i*. *LoanCH* represents loan characteristics and is a vector of two variables (i.e., loan amount and loan term). Similarly, *BorrowerCH* represents borrower characteristics and is a vector of four variables (i.e., annual income, home ownership, DTI ratio, and revolving balance). The *LoanCH* and *BorrowerCH* vectors are assumed to be exogenous in the model. Finally, *Controls* is a vector of dummy variables for loan purposes that is used to control for each loan's inherent risk.

Equation (1) is the structural equation of our main interest, and Equation (2), the first-stage equation, identifies instruments for the endogenous regressor (i.e., loan interest rate) with exogenous variables and control variables. Because we used a total of four instruments for one endogenous variable in the model, the order condition for identification was satisfied (Wooldridge, 2015). By assumption, $(u_i, v_i) \sim N(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix

$$\text{Var}(u_i, v_i) = \Sigma = \begin{bmatrix} 1 & \sigma'_{21} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \quad (3)$$

where σ_{11} is normalized to one to identify the model. The multivariate normality of (u_i, v_i) implies that $u_i|v_i = \rho v_i + \varepsilon_i$, where $E(\varepsilon_i|v_i) = 0$ (Cameron & Trivedi, 2010). Thus, a test of H_0: $\rho = 0$ is equivalent to a test of exogeneity of our endogenous variable, *IntRate* (see Table 3).

Results

Table 3 shows the parameter estimates of the first-stage IV regression (i.e., Equation 2) and the second-stage probit model (i.e., Equation 1). We found that our instruments, as well as other exogenous variables, explained most of the variation in loan default risk. In the first-stage IV regression, the estimates of all IVs were highly significant and, thus, qualified as instruments. The loan interest rate was shown to decrease with a positive and detailed loan description, and higher borrower grades and fewer previous delinquency records were also significantly associated with low interest rates. This certainly makes sense given that the P2P lending company adjusts the interest rate based on the borrower's overall grade, including credit history. Our results also suggest that a loan description can be one of the important factors for P2P loan companies and lenders to consider when evaluating the loan.

The test result for exogeneity indicates that *IntRate* is an endogenous regressor. The null-hypothesis of exogeneity, (i.e., $H_0 : \rho = 0$), was rejected at the 0.05 level, and the estimated coefficient of ρ was significantly positive. This implies that unmeasured factors increasing a loan interest rate also make a loan more likely to default, conditional on other regressors included in the model.

The estimates of the second stage-probit model, which is our main interest, suggest that loan default risk is highly dependent on loan characteristics, borrower characteristics, and loan purposes. We found that all the loan characteristics had significant positive effects on loan default risk. That is, risky loans with longer terms, larger amounts, and high interest rates were more likely to default. In terms of borrower characteristics, our results suggest that the probability of loan default is also relevant to the borrower's financial situation. For high-income borrowers, the likelihood of loan default was lower. Other measures of the borrower's financial status, such as the DTI ratio and revolving balance, were also shown to influence loan default risk. It turned out that default risk increased as the proportion of DTI increased, while a high revolving balance lowered the loan default rate. Because the DTI ratio is defined as the percentage of a person's overall income that goes toward paying debts, loans requested by borrowers with a low DTI ratio are more likely to exhibit low default rates. On the other hand, because a revolving balance is defined as a total balance owing on credit cards, it can be related to the borrower's general capacity to spend. Borrowers whose revolving balances are high may have greater purchasing power and pay more attention to their economic costs of default, such as a credit score drop. A recent study based on the two largest U.S. P2P lenders, *LendingClub* and *Prosper*, also reported that loans with high revolving credit balance performed better on return on investment (Renton, 2012).

As a control vector, the loan purpose was shown to be another significant factor influencing the default rate. We found that loans for small businesses were the riskiest ones, which was compatible with previous research (see Polena & Regner, 2016; Serrano-Cinca et al., 2015). Wedding or

TABLE 3. Estimates of a Binary Probit Model With IV

	First Stage(Dependent Variable: IntRate)		Second Stage(Dependent Variable: loan default)	
Variables	**Estimates**	**Std. Err.**	**Estimates**	**Std. Err.**
Instruments (IVs)				
Sentiment score	−0.066***	0.011		
Total words	−0.002***	0.000		
Grade	−3.267***	0.002		
Delinquent history	0.062***	0.004		
Loan characteristics				
Interest rate			0.060***	0.001
Loan amount	−0.026***	0.004	0.050***	0.006
Loan term	0.217***	0.008	0.141***	0.010
Borrower characteristics				
Annual income	−0.008***	0.001	−0.028***	0.001
Home ownership	0.018***	0.010	−0.012**	0.013
DTI ratio	0.012***	0.000	0.008***	0.001
Revolving balance	−0.007***	0.002	−0.010***	0.003
Loan purposes (Dummies)				
Car	−0.371***	0.030	−0.045**	0.042
Credit card	−0.096***	0.014	−0.068***	0.018
Debt consolidation	−0.049***	0.013	0.003**	0.017
Home improvement	−0.191***	0.017	0.021**	0.023
House	−0.217***	0.038	0.064**	0.050
Major purchase	−0.262***	0.023	−0.024*	0.032
Medical	−0.009**	0.033	0.021**	0.042
Moving	0.030**	0.039	−0.033*	0.050
Renewable energy	−0.068***	0.108	0.111**	0.134
Small business	−0.139***	0.026	0.305***	0.031
Vacation	0.093**	0.041	−0.007**	0.053
Wedding	−0.074**	0.035	−0.132***	0.047
Others (omitted)				
Constant	31.339***	0.019	−1.948***	0.025
Endogeneity (ρ)			0.009**	0.004
Variance (σ_{22})			1.395***	0.002
Wald test of exogeneity: $\chi^2(1) = 5.86$, Prob > $\chi^2 = 0.0155$				

Note. DTI = debt-to-income; IV =instrumental variable.
***, **, and * represent significance at the 1%, 5%, 10% levels, respectively. ρ is the correlation coefficient between u_i and v_i in our main equations. A test for exogeneity of *IntRate* is implemented as a Wald test of $\rho = 0$ using the usual standard errors. Among the dummies for loan purposes, we omit one dummy (i.e., loans for *other* purposes) for a base group.

credit card loans were shown to be less risky compared to other loans. This finding is also consistent with the studies of Serrano-Cinca et al. (2015), suggesting that a well-categorized loan purpose is an important factor for lenders to consider before investing. Overall, the model estimation results support our hypothesis that P2P loan characteristics and borrower characteristics are associated with both loan interest rates and loan default risk (i.e., support hypothesis 1).

Finally, we calculated the average marginal effects by using our probit model with an endogenous regressor. Since a probit model is non-linear, the estimated coefficients are often difficult to interpret directly. In probit models, the marginal effect of changing a variable is not constant, and it depends on the value of other variables and their coefficients. Thus, the marginal effect of the j^{th} regressor on the probability of loan i's default can be expressed as

$$\frac{\partial \Pr\left[y_i = 1 | \mathbf{x}_i\right]}{\partial \mathbf{x}_{ij}} = (\mathbf{x}_i' \beta)\, \beta_j \qquad (4)$$

where $\mathbf{x}_i$ is a regressor vector *except* the j^{th} regressor, and $\phi(\mathbf{x}'_i\beta)$ represents the value of the standard normal probability density function at $\mathbf{x}'_i\beta$. The equation shows that the marginal effect is determined not only by β_j, but also by the values of all other variables in the model. Therefore, the marginal effect will differ for each loan due to its distinct loan characteristics.

Table 4 shows the marginal effect of each variable in our probit model, averaging across 188,181 loan cases. We see, for example, that a one unit increase in *IntRate* leads to an average increase of 0.013 in the probability of default risk.

TABLE 4. Average Marginal Effects by Using the IV Probit Model

Variables	Estimates (dy/dx)	Std. Err.
Loan characteristics		
Interest rate	0.013***	0.000
Loan amount	0.011***	0.001
Loan term	0.032***	0.002
Borrower characteristics		
Annual income	–0.006***	0.000
Home ownership	–0.003***	0.003
DTI ratio	0.002***	0.000
Revolving balance	–0.002***	0.001
Loan purposes (Dummies)		
Car	–0.010**	0.009
Credit card	–0.015***	0.004
Debt consolidation	0.001**	0.004
Home improvement	0.005**	0.005
House	0.014**	0.011
Major purchase	–0.005**	0.007
Medical	0.005**	0.009
Moving	–0.007**	0.011
Renewable energy	0.025**	0.030
Small business	0.068***	0.007
Vacation	–0.002**	0.012
Wedding	–0.029***	0.010
Others (omitted)		

Note. DTI = debt-to-income; IV =instrumental variable. ***, **, and * represent significance at the 1%, 5%, 10% levels, respectively.

Conclusion

The goal of our study was to develop and empirically test a model to understand P2P loan default risk. Our research emphasizes the importance of considering the sentiment of unstructured text and endogenous effects in modeling default risk in P2P lending. Because unmeasured or omitted variables both affect other covariates and the dependent variable (i.e., the probability of default) simultaneously, we focused on the loan interest rate which is determined by a P2P lending platform and control for endogeneity through the IV method. Our main findings were that small short-term loans with low interest rates are less likely to default, and the borrower's good financial situation is shown to be a critical factor in successful loan repayment. Furthermore, loan default rates were found to differ depending on the loan purpose, and P2P loans required for operating small businesses had relatively higher default rates in particular.

In terms of the loan description, we found that a loan described positively and specifically with many words was favored to receive borrower-favorable loan conditions with low interest rates. Writing a positive and detailed loan description can help P2P lending companies understand the requested loan and adjust the interest rate accordingly. Our study suggests that a favorable loan description followed by low interest rates is not only beneficial to borrowers, but also important for lenders in determining whether to invest in the loan. This finding implies that P2P loan decisions are affected by subjective perceptions about borrowers inferred from their personal information (Duarte, Siegel, & Young, 2012). P2P lending companies and financial consultants should put forth more effort to ensure that accurate and reliable information is provided to lenders because the possibility of information being distorted or biased can be increased due to the characteristics of the online platform. One fruitful avenue for future research, therefore, would be to examine whether online P2P lending service users are more likely to make biased decisions compared with when

they go through the same loan application and qualification process for offline-based financial institutions.

Our findings also suggest that P2P loan interest rates are affected by not only the sentiment reflected in the loan description, but also the financial status and credit information of the borrower, applied loan amount, and the loan period. That is, borrowers with less severe past delinquencies and better credit ratings tended to enjoy lower interest rates, while those with lower income or longer loan periods were more likely to experience higher interest rates. Financial planners will be able to develop adequate loan assessment criteria by focusing on the association between the P2P loan default rates and interest rates as well as by considering factors that can affect these elements.

Our research contributes to P2P lenders and risk managers who are interested in planning their own optimal investment strategies. Using our model and results, we expect that investors can compute a default risk for a loan and diversify their P2P lending portfolios more effectively. A key part of portfolio management is finding the right mix of different assets in the portfolio. Thus, improving default risk prediction will become increasingly important, and we expect our research will serve as essential input for portfolio optimization, diversification, and risk management. Future research will benefit from formulating portfolio optimization in P2P lending using both structured and unstructured data.

Limitations and Future Directions

Our study goes beyond the perspective of topics and methodologies employed by previous research on the P2P lending market, making new attempts in terms of data and modeling that contribute to the expansion of research efforts associated with P2P loan default risks. Despite contributions such as the implementation of a new type of data (i.e., unstructured data) and the proposal of a new model (i.e., an IV probit model) that reduces bias in the estimation methods used by existing models, our study reveals limitations caused by difficulties in collecting the additional data required for more precise analysis.

To predict the loan default risk using an econometric model, we extracted model variables from secondary data provided by *Lending Club*, a P2P lending company. Moreover, this study considered the characteristics of borrowers that affect the loan default risk, such as annual income, home ownership, DTI ratio, and revolving balance. However, personal characteristics such as education level, gender, occupation, and prior experience and familiarity with the P2P loan market of the borrower can also influence the loan default rate (Chen et al., 2018; Han et al., 2019). Furthermore, the secondary data employed by the current study only included measurable quantitative data about the borrowers and the history of loan products to which they previously applied, but future research will need to collect additional data through survey and other methods to account for psychological factors indicating borrower personalities, which may also affect the P2P loan default risk. Such a measure is important in that it expands the range of potential IVs that can be considered in the loan default prediction model suggested by the current study.

Over and above the quantified structured data employed in many previous studies, our study utilized unstructured data with a text format including a loan description as a factor that can influence P2P loan default risk. However, future research needs to utilize more diverse qualitative data. P2P loan borrowers may have different attitudes to credit management depending on their personal situations, which may include psychological factors that are not reflected in the secondary data. In fact, Szendrey and Fiala (2018) conducted an online survey to demonstrate that the borrowers engage in different credit management behaviors depending on their self-perceived future economic status and income level. Such unobserved mechanisms based on economic, psychological, and environmental factors may exist, and to reveal them, we need to conduct big data analysis with more variables or additional research that seeks causality between explanatory variables and dependent variables through a well-designed experiment. Research on the borrowers' financial behaviors or attitude towards P2P loans will help predict P2P loan performance; based on such findings, financial planners will be able to provide more effective consulting services to P2P loan investors.

Lastly, the current study is limited in terms of the generalizability of its results. Although we analyzed data from *Lending Club*, a major company in the U.S. P2P loan market, P2P lending behaviors of customers in other major P2P lending markets such as Europe and China may differ. As

such, a potentially important research topic in the future may include a comparative analysis of factors influencing the loan default risk in different countries' P2P loan markets and examining government policies or cultural backgrounds that can explain differences. In addition, because the customer base participating in the P2P loan market is diversifying to include institutional investors, individual businesses, and corporations, a new P2P loan default prediction model that takes the heterogeneity of such market participants into account may need to be developed.

Implications

Based on the structured and unstructured data, our study investigated the characteristics of borrowers and loans that may affect the P2P loan default risk. The analysis revealed that P2P loan default risk may be mitigated by the loan policies or guidelines of the P2P lending companies. That is, the most important thing for the P2P loan service providers is to encourage borrowers to apply for appropriate loans to their financial situations. This will regulate interest rates that borrowers need to pay in the future properly, and ultimately contribute to decreasing the likelihood of loan default. Personalized financial advice must be provided based on data analysis, which may be achieved by utilizing the model suggested in the current study that measures the loan default risk: The model allows financial counselors to quantitatively explain the degree of effect that each model variable may have on the loan interests rates and default risk. Furthermore, this model can help financial counselors assist the decision-making process of both investors and borrowers who may be less familiar with the P2P loan market, in addition to enhancing the likelihood of loan repayment and reducing the default risk, thereby inducing desirable financial behaviors (Moreland, 2018).

Regarding the data required to review P2P loan applications, financial planners (who manage or advise P2P lending services) need to consider unstructured data about loan applicants (e.g., social network behavior) to enable more accurate review and default prediction. Our study demonstrated that the loan default risk can be predicted through not only the conventionally used quantified structured data such as loan amount, annual income, and interest rates, but also sentiment scores extracted from unstructured data such as text-based loan descriptions. Recent developments in the financial industry saw the emergence of various fintech companies based on artificial intelligence technologies. This enabled financial institutions to fortify the conventional financial transaction data used for individual credit assessments with non-financial information including social media activities (e.g., social media posting topics) and online reputation of an individual (e.g., number of social media friends). For example, to predict the probability that an applicant will repay the loan, *ZestFinance*, a fintech company in the United States, breaks away from the traditional credit assessment method based on a small number of variables employed by commercial banks, and processes more than three thousand different variables in an independently-developed machine learning model (ZestFinance Blog, 2019). ZestFinance is working to enhance not only the accuracy of this loan application review process, but also its fairness and explainability of the results (Fuscaldo, 2019). This means that while it is important to accurately predict the default probability of a loan, financial counselors must also be able to understand the loan application review model and explain which variables pertaining to the applicant led to the approval or rejection of the loan application. Understanding the data-based loan application assessment model mechanism can be useful for financial counselors when they explain the risks of the P2P loan to investors.

Unlike in the case of structured data, implementing unstructured data in a prediction model to obtain reliable results requires a significant amount of data cleansing work. As indicated by the maxim, "garbage in, garbage out" (Kilkenny & Robinson, 2018), unrefined data with poor quality degrades the reliability of analysis results. Advances in technologies for refining and analyzing unstructured data, such as the sentiment analysis method introduced in the current study, can provide new opportunities to the loan consumers who were isolated from the loan market centered around financial institutions. Conventional financial institutions and credit rating agencies basically used to rely on numerical financial transaction data. As such, younger adults, homemakers, and freelancers who do not have a bank account or credit history through loans and other transactions were disadvantaged under this previous loan application assessment method based on financial transaction history. Low credit ratings hinder consumers from using loan services, and in turn generate further lack of credit transaction history, thereby creating a vicious cycle. Future advances in artificial intelligence and other technologies enhancing big data analysis algorithms, as well

as the increase in the utilization of unstructured data that show personal psychologies or life patterns, will facilitate the financial consumers' access to loan services. In addition, advances in prediction models that simultaneously consider numerous variables will enable more precise credit analysis, allowing customers who were not able to receive a proper credit rating to procure loan services at lower interest rates. Conversely, investors will also be able to increase their profit margins because technological advances will enable them to invest in relatively less risky loan products. As such, financial counselors will need to take note of the growth in the P2P loan market based on data analysis and identify niche market opportunities that were not formed in the credit lending market until recently. Furthermore, they will need to plan and develop services that enable more small loans to be safely offered at reasonable interest rates.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. https://doi.org/10.1080/01621459.1996.10476902

Berger, S. C., & Gleisner, F. (2009). Emergence of financial intermediaries in electronic markets: The case of online P2P lending. *Business Research*, *2*(1), 39–65. https://doi.org/10.1007/BF03343528

Beutler, T., Bichsel, R., Bruhin, A., & Danton, J. (2020). The impact of interest rate risk on bank lending. *Journal of Banking & Finance, 115,* 105797. https://doi.org/10.1016/j.jbankfin.2020.105797.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York, NY: Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2010). *Microeconometrics using stata: Revised edition*. College Station, TX: Stata Press

Chen, H., De, P., Hu, Y., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, *27*(5), 1367–1403. https://doi.org/10.1093/rfs/hhu001

Chen, J., Jiang, J., & Liu, Y. J. (2018). Financial literacy and gender difference in loan performance. *Journal of Empirical Finance*, *48*, 307–320. https://doi.org/10.1016/j.jempfin.2018.06.004

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, *29*(5), 944–957. https://doi.org/10.1287/mksc.1100.0572

Dietrich, A. (2016). What drives the interest rates in the P2P consumer lending market? Empirical evidence from Switzerland. World Finance Conference, New York.

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies*, *25*(8), 2455–2484. https://doi.org/10.1093/rfs/hhs071

Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2014). Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. *Applied Economics*, *47*(1), 54–70. https://doi.org/10.1080/00036846.2014.962222

Fuscaldo, D. (2019, March 19). *ZestFinance using AI to bring fairness to mortgage lending*. Retrieved from https://www.forbes.com/sites/donnafuscaldo/2019/03/19/zestfinance-using-ai-to-bring-fairness-to-mortgage-lending/#4ff53ad47f2d

Garret, O. (2017). *How to join the P2P lending revolution and earn +10% yields*. Retrieved from https://www.forbes.com/sites/oliviergarret/2017/03/06/how-to-join-the-p2p-lending-revolution-and-earn-10-yields/2/#19e0faf7662e

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, *249*(2), 417–426. https://doi.org/10.1016/j.ejor.2015.05.050

Han, L., Xiao, J. J., & Su, Z. (2019). Financing knowledge, risk attitude and P2P borrowing in China. *International Journal of Consumer Studies*, *43*(2), 166–177. https://doi.org/10.1111/ijcs.12494

Heber, A. (2015). *MORGAN STANLEY: Peer-to-peer lending is about to really take off in Australia*. Retrieved from https://www.businessinsider.com.au/morgan-stanley-peer-to-peer-lending-is-about-to-really-take-off-in-australia-2015-5

Hendricks, N. P., Janzen, J. P., & Smith, A. (2015). Futures prices in supply analysis: Are instrumental variables necessary? *American Journal of Agricultural Economics*, *97*(1), 22–39. https://doi.org/10.1093/ajae/aau062

Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications*, *41*(4–1), 1041–1050. https://doi.org/10.1016/j.eswa.2013.07.101

Kilkenny, M., & Robinson, K. (2018). Data quality: "Garbage in –Garbage out". *Health Information Management Journal*, *47*(3), 103–105. https://doi.org/10.1177/1833358318774357

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: Evidence from a lending platform in China. *Applied Economics*, *49*(35), 3538–3545. https://doi.org/10.1080/00036846.2016.1262526

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Mild, A., Waitz, M., & Wöckl, J. (2015). How long can you go? –Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. *Journal of Business Research*, *68*(6), 1291–1305. https://doi.org/10.1016/j.jbusres.2014.11.021

Moreland, K. A. (2018). Seeking financial advice and other desirable financial behaviors. *Journal of Financial Counseling and Planning*, *29*(2), 198–207. https://doi.org/10.1891/1052-3073.29.2.198

Nowak, A., Ross, A., & Yencha, C. (2018). Small business borrowing and peer-to-peer lending: Evidence from lending club. *Contemporary Economic Policy*, *36*(2), 318–336. https://doi.org/10.1111/coep.12252

Petrin, A., & Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, *47*(1), 3–13. https://doi.org/10.1509/jmkr.47.1.3

Pokorná, M., & Sponer, M. (2016). Social lending and its risks. *Procedia Social and Behavioral Sciences*, *220*(31), 330–337. https://doi.org/10.1016/j.sbspro.2016.05.506

Polena, M., & Regner, T. (2016). *Determinants of borrowers' default in P2P lending under consideration of the loan risk class*. Jena Economic Research Papers, No. 2016-023, Friedrich Schiller University Jena, Jena.

Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: A review. *Journal of Research in Interactive Marketing*, *12*(2), 146–163. https://doi.org/10.1108/JRIM-05-2017-0030

Renton, P. (2012). *Three myths investors have about P2P lending*. Retrieved from https://www.lendacademy.com/three-myths-investors-have-about-p2p-lending/

Rutz, O. J., & Watson, G. F. IV. (2019). Endogeneity and marketing strategy research: An overview. *Journal of the Academy of Marketing Science*, *47*(3), 479–498. https://doi.org/10.1007/s11747-019-00630-4

Schweidel, D., & Moe, W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *Journal of Marketing Research*, *51*(4), 387–402. https://doi.org/10.1509/jmr.12.0424

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, *89*, 113–122. https://doi.org/10.1016/j.dss.2016.06.014

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLOS ONE*, (). doi:https://doi.org/10.1371/journal.pone.0139427

Shin, S. H., & Kim, K. T. (2018). Income uncertainty and household stock ownership during the great recession. *Journal of Financial Counseling and Planning*, *29*(2), 383–395. https://doi.org/10.1891/1052-3073.29.2.383

Szendrey, J., & Fiala, L. (2018). "I think I can get ahead!" Perceived economic mobility, income, and financial behaviors of young adults. *Journal of Financial Counseling and Planning*, *29*(2), 290–303. https://doi.org/10.1891/1052-3073.29.2.290

Tang, H. (2019). Peer-to-peer lenders versus banks: Substitutes or complement? *The Review of Financial Studies*, *32*(5), 1900–1938. https://doi.org/10.1093/rfs/hhy137

Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, *31*(2), 198–215. https://doi.org/10.1287/mksc.1110.0682

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach* (6th ed.). Independence, KY: Cengage.

ZestFinance Blog. (2019, April 9). *Using ML to increase your loan portfolio yield*. Retrieved from https://www.zestfinance.com/blog/using-ml-to-increase-loan-portfolio-yield

Disclosure. This study was conducted by research funds from Gwangju University in 2020.

Acknowledgement. This study was conducted by research funds from Gwangju University in 2020.

Funding. The author(s) received no specific grant or financial support for the research, authorship, and/or publication of this article.