

When Words Sweat:

Identifying Signals for Loan Default in the Text of Loan Applications

The authors present empirical evidence that borrowers, consciously or not, leave traces of their intentions, circumstances, and personality traits in the text they write when applying for a loan. This textual information has a substantial and significant ability to predict whether borrowers will pay back the loan over and beyond the financial and demographic variables commonly used in models predicting default. The authors use text-mining and machine-learning tools to automatically process and analyze the raw text in over 18,000 loan requests from Prosper.com, an online crowdfunding platform. The authors find that loan requests written by defaulting borrowers are more likely to include words related to their family, mentions of god, short-term focused words, the borrower's financial and general hardship, and pleading lenders for help. The authors further observe that defaulting loan requests are often written in a manner consistent with the writing style of extroverts and liars.

INTRODUCTION

Imagine you consider lending \$2,000 to one of two borrowers on a crowdfunding website. The borrowers are identical in terms of their demographic and financial characteristics (e.g., credit score), the amount of money they wish to borrow, and the reason for borrowing the money. However, the text they provided when applying for a loan differs: Borrower #1 writes “I am a hard working person, married for 25 years, and have two wonderful boys. Please let me explain why I need help. I would use the \$2,000 loan to fix our roof. Thank you, god bless you, and I promise to pay you back.” while borrower #2 writes “While the past year in our new place has been more than great, the roof is now leaking and I need to borrow \$2,000 to cover the cost of the repair. I pay all bills (e.g., car loans, cable, utilities) on time.” Who is more likely to default on her loan? This question is at the center of our research, as we investigate the power of words in predicting loan default. As we discuss later the text and writing style of borrower #1 include many traces commonly found in defaulting loan requests. In fact, all else equal, our analyses shows that based on the loan request text, borrower #1 is approximately eight times more likely to default relative to borrower #2.

Unprecedented loan default levels were at the heart of the financial crisis of 2008-9 (Lewis 2015). While the majority of these loans were mortgages, a significant amount was in consumer loans. As a result, it became difficult for consumers to secure a bank loan—as much as 30% of the people who wished to get a loan were not able to obtain one through conventional channels (Dogra and Gorbachev 2016). Consequently, Americans turned to other sources. In 2011, 14% of American households (over 16 million households) reported using “non-bank credit,” which includes payday loans and pawn shops (Mills and Monson 2013). Arguably, a better alternative to these “shark” loans emerged in the form of online crowdfunding platforms,

which serve as a marketplace for lenders and borrowers. These platforms have become a substantial financial market, with over 1,250 sites and \$34.4 billion in raised funds in 2015 (Crowdfunding industry report, Massolution 2015).

While the idea of crowdfunding may be appealing to both borrowers, who may not be able to get a loan elsewhere, and lenders, who may be able to get better returns on their investments, default rates in crowdfunding sites are high, which put the entire concept at risk. In comparison to more secured channels such as bank loans, default rates are especially important in crowdfunded loans because the crowdfunding process is riskier and more uncertain—lending money to strangers without any collateral. Furthermore, the online nature of crowdfunding platforms eliminates the human interactions around the financial transaction of granting a loan, which has been shown to reduce default rates. Indeed, Agarwal and Hauswald (2010) found that supplementing the loan application process with the human touch of loan officers decreases default rate significantly due to better screening on the bank's part and higher interpersonal commitment on the consumer's part. We propose that in a similar vein to loan officers, who are able to assess default likelihood in the offline banking environment, the text that borrowers write when requesting an online crowdfunded loan may leave some traces—similar to body language detected by loan officers—that can improve predictions of future repayment behavior. While it is known that our demeanor can be a manifestation of our true intentions (DePaulo et al. 2003), it is not obvious whether this idea transfers to the text we write and particularly in the context of online loan applications, which should be rational and purposeful.

The objective of the current research is to investigate whether the text borrowers write when applying for a crowdfunded loan can be predictive of loan default. Specifically, we investigate whether borrowers leave traces of their intentions, circumstances, emotional states,

and personality in the text they write that are predictive of whether they will default on their loan up to three years after the text was written. To answer these questions we apply text-mining and machine-learning tools to a dataset of over eighteen thousand loans from the crowdfunding platform Prosper.

Predicting loan default is often a difficult task because loans are repaid over a lengthy period of time, during which unforeseen circumstances may arise. For that reason, traditional lenders (such as banks) and researchers have focused on collecting and processing as many pieces of information as possible within this tightly regulated industry. Most telling is, of course, the borrower's financial strength (Avery, Calem, and Canner 2004), which is manifested by credit history and FICO scores, income, and debt (Mayer, Pence, and Sherlund 2009; Thomas 2000). Demographics, such as race, gender, and geographic location, have also been shown to correlate with repayment (Rugh and Massey 2010). In addition, loan characteristics such as the amount that is being borrowed and the interest rate may attribute to the probability of its repayment (Gross, Cekic, Hossler, and Hillman 2009). These factors have also been found to predict loan granting in crowdfunding platforms (Herzenstein et al. 2008; Berger and Gleisner 2009; Zhang and Liu 2012).

More recently research on crowdfunded loans has supplemented the traditional focus on financials and demographics metrics with new sources of data, such as using pictures of borrowers to rate their attractiveness (Duarte, Siegel, and Young 2012; Pope and Sydnor 2011; Ravina 2012), using the extent of herding to fund a loan auction to learn about the worthiness of a borrower (Herzenstein, Dholakia, and Andrews 2010; Luo et al. 2011), and using the narrative crafted by the borrower (Herzenstein, Sonenshein, and Dholakia 2011; Iyer et al. 2009; Michels 2012; Sonenshein, Herzenstein, and Dholakia 2011).

We add to this literature by investigating the value of words in the borrowers' loan request in indicating their likelihood of repaying the loan. Our approach differs from the aforementioned papers in two important ways. First, we automatically text-mine the raw text that borrowers write in their loan request and analyze the impact of the overall writing style, as well as individual words on likelihood of default, while controlling for the effect of traditional measures, such as credit score and demographics. We do so using text-mining and machine learning tools, which allows us to scale up our analysis to many thousands of loan requests and a large corpora of text (millions of words). Consequently, our analysis is more inclusive—we examine all words and themes that emerge from the data, rather than focus on a small subset that was pre-coded. Second, unlike previous papers we attempt to *predict* which borrowers are more likely to default based on the text they write (along with traditional measures).

To investigate the value of the loan application text in predicting default we created an ensemble of predictive models consisting of decision trees and regularized logit models. We find that the predictive ability of the textual information alone is of similar magnitude to that of the financial and demographic information. Moreover, supplementing the financial and demographical information with the textual information improves predictions of default by as much as 4.03%.

Next we use a multi-method approach to uncover the words and writing styles that are most predictive of default. Using a naïve Bayes and an L1 regularization binary logistic model we find that loan requests written by defaulting borrowers are more likely to include words related to the borrower's family, financial and general hardship, mentions of god and the near future, as well as pleading lenders for help. We use a latent Dirichlet allocation (LDA) to identify the loan purpose, life circumstances, and writing styles that are most associated with loan

default. We find that loans whose purpose is to help with a business or medical circumstances are riskier than other types loans in terms of their default likelihood. Pleading lenders for help and providing explanations are also associated with higher risk of default, consistent with the naïve Bayes results.

We further explore the writing styles and personality traces embedded in the loan request text using the Linguistic Inquiry and Word Count dictionary (LIWC; Tausczik and Pennebaker 2010). We find that defaulting loan requests are written in a manner consistent with the writing style of extroverts and liars. While we are unable to claim that defaulting borrowers were intentionally deceptive when they wrote the loan request, we believe their writing style may have reflected their doubts in their ability to repay the loan.

The rest of this paper is organized as follows. In the next section we delineate the data and our text-mining approach. We then describe the ensemble stacking approach of decision trees and regularized logistic regressions, which we use to assess the ability of the textual information to predict default. In the following section we use a combination of approaches including a naïve Bayes, an L1 regularization binary logistic model, an LDA analysis, and the LIWC dictionary to investigate which words and writing styles are most likely to appear in defaulting loans. We discuss our results vis-à-vis extant literature from linguistic, psychology, and economics.

SETTINGS AND DATA

We examine the predictive power of text on default using data from Prosper.com, the first online crowdfunding platform and currently the second largest in the United States, with over 2 million members and \$7 billion in funded unsecured loans. In prosper, potential borrowers

submit their request for a loan for a specific amount with a specific maximum interest rate they are willing to pay, and lender then bid in a Dutch-like auction on the lender rate for loan.¹ We downloaded all loan requests posted between April 2007 and October 2008, a total of 137,952 listings. This time frame is part of “Prosper 1.0”. In October 2008, the Securities and Exchange Commission required Prosper to register as a seller of investment, and when Prosper re-launched in July 2009 it made significant changes to the platform, thus named “Prosper 2.0”. We chose data from “Prosper 1.0” because it is richer and more diverse (particularly with respect to the textual information in the loan request) due to stricter guidelines imposed on borrowers in “Prosper 2.0”.

When posting a loan request on Prosper potential borrowers have to specify the loan amount they wish to borrow (between \$1,000 and \$25,000 in our data), the maximum interest rate they are willing to pay, and other personal information, such as debt to income ratio and whether they are home owners. Prosper verifies all financial information including the potential borrower’s credit score from Experian, and assigns each borrower a credit grade that reflects all of this information. The possible credit grades are AA (lowest risk for lenders), A, B, C, D, E, and HR (highest risk to lenders). See correspondence between Prosper’s credit grades and FICO score ranges in Table A1 in the Web Appendix. In addition, borrowers can upload as many pictures as they wish, and use an open textbox to write any information they wish, with no length restriction. The words borrowers use in that textbox are at the center of our research.

Because we are interested in predicting default, we focus on those loan requests that were funded, 19,446 requests. Our final dataset contains funded loan requests that have some text in

¹ In 2009 Prosper cancelled the bidding process and moved to an interest rate that is calculated by Prosper. Our dataset precedes this change. Because our focus is on loan default given that the loan was granted, we are not modeling the bidding process, but we control in our analyses for the interest rate that results from this bidding process.

the textbox—18,312 loans. The default rate in our sample of loans is 33.1%.²

Text Mining

We automatically text-mined the raw text in each loan application using the *tm* package in R. Our textual unit is a loan application. For each loan application, we first tokenize each word, a process that breaks down each loan application into the distinct words it contains. We then use Porter's stemming algorithm, to collapse variations of words into one. For example, "borrower," "borrowed," "borrowing," and "borrowers" become "borrow". In total, the loan requests in our dataset have over 3.5 million words, corresponding to 30,920 unique words that are at least 3 letters long (we also excluded from our analysis numbers and symbols).³ In addition to words/stems we also look at two-word combinations (an approach often referred to as n-gram, in which for $n = 2$, we get bi-grams). To reduce the dimensionality of the textual data and avoid biasing our results toward more obscure words, we focus our analyses on the most frequent stemmed words and bi-grams that appeared in at least 400 loan requests. We are left with 1,032 bi-grams.⁴

Textual, Financial, and Demographic Variables

Our dependent variable is loan default as reported by Prosper⁵ (binary: 1 = paid in full, 0 = defaulted). Because our data horizon ends in 2008, and all Prosper loans at the time were to be repaid over three years or less, we know whether each loan in our database was repaid or

² The default rate in the loans with no text that were dropped is 35%, which is similar to the default rate in our sample.

³ Because of stemming, words with less than 3 words such as "I" may be kept due to longer stems (e.g., I've).

⁴ We checked the robustness of our analyses to increasing the number of words and bi-grams included in the analysis. Our results did not change qualitatively when we increased the number of bi-grams.

⁵ We classified a loan as "defaulted" if the loan status in Prosper is "Charge-off," "Defaulted (Bankruptcy)," or "Defaulted (Delinquency)." We classified a loan as "paid" if it is labeled "Paid in full," "Settled in full," or "Paid."

defaulted. Our independent variables include textual, financial, and demographic variables.

Textual Variables

The textual variables include: (1) The *number of characters* in the title and the textbox in the loan request. The length of the text has been associated with deception, with longer texts more likely to be written by liars. Hancock et al. (2007) showed that liars wrote much more when communicating via text messages than non-liars. Similarly, Ott, Cardie, and Hancock (2012) demonstrated that fake hospitality reviews are wordier though less descriptive. However, in the context of online dating websites, Toma and Hancock (2012) showed that shorter profiles are indicative the person is lying, because they wished to avoid certain topics. (2) The *percent of words with six or more letters*. This metric is commonly used to measure complex language, education level, and social status (Tausczik and Pennebaker 2010). More educated people are likely to have higher income and higher levels of financial literacy and hence are less likely to default on their loan, relative to less educated people (Nyhus and Webley 2001). But the use of complex language can also be risky if readers of the text perceive it to be artificially or frivolously complex. Indeed, Oppenheimer (2006) demonstrated, in the context of admission essays, that if complex vocabulary is used superfluously, the author may face a detrimental outcome, suggesting the higher language was likely used deceptively. (3) The *Simple Measure of Gobbledygook* (SMOG; McLaughlin, 1969), which measures writing quality by mapping it to number of years of formal education needed to easily understand the text in first reading. (4) A count of *spelling mistakes* based on the enchant spell checker using the Pyenchant 1.6.6. package in Python. Harkness (2016) shows that spelling mistakes are associated with a lower likelihood of granting a loan in traditional channels. (6) The *bi-grams* from the open textbox in each loan application following the text mining process described earlier.

Because loan requests differ in length, and words differ in the frequency of appearance in our corpus, we normalize the frequency of a word appearance in a loan request to its appearance in the corpus and the number of words in the loan request using the term frequency–inverse document frequency, tf-idf, measure commonly used in information retrieval. The term frequency for word j in loan request m is defined by $tf_{jm} = X_{jm}/N_m$, where X_{jm} is the number of times word j appeared in loan request m , and N_m is the number of words in loan request m . The inverse-document-frequency is defined by $idf_j = \log(D/M_j)$, where D is the number of loan requests and M_j is the number of loan requests in which word j appeared. *Tf-idf* is given by $tf - idf_{jm} = tf_{jm} \times (idf_j + 1)$.

Financial and Demographic Variables

The second type of variables we consider are financial and demographic information, commonly used in traditional risk models. We attempt to control for all information available to lenders on Prosper, including the loan amount, borrower’s credit grade (modeled as a categorical variable AA-HR), debt to income ratio, whether the borrower is a home owner, the bank fee for payment transfers, whether the loan is a relisting of a previous unsuccessful loan request, and whether the borrower included a picture with the loan.

In order to truly account for all the information lenders have when viewing a loan request, we extracted information included in the borrowers profile pictures, such as gender, age bracket, and race using human coders. About a third of the borrowers’ profiles in our data (6,078 profiles) included at least one picture that is not a stock photo, however many pictures were not of the borrower, or included more than one person. To identify the borrower in the picture we manually coded the borrower’s profile pictures, using the following process. If the picture included captions, we relied on it to identify the borrower (for example, “My lovely wife and I”).

If the picture did not include captions and there was one adult in the picture, we assumed the adult in the picture was the borrower (following the procedure in Pope and Sydnor 2011). Once borrowers were identified, we recorded their gender (female, male, “cannot tell”), age (in three brackets: young, middle-aged, old), and race (Caucasian, African American, Asian, Hispanic, or “cannot tell”). If the picture included more than one adult and there were no captions or if the picture did not include any adult (e.g., the picture included kids, pets, or a kitchen project) we could not identify the borrower and therefore defined the gender and race of that picture as undefined. We augmented the age in unidentified pictures with the average age of the identified pictures with the three ages categories coded as 1, 2 and 3, respectively.

Each picture was evaluated by at least two different undergraduate student coders, who were unaware of the research objective. Cohen Kappas suggest fairly high levels of agreement across coders, gender = 0.89, race = 0.67, and age = 0.44.⁶ Disagreements were resolved by an additional coder who served as the final judge, observing the rating of the previous coders.⁷

In addition to the aforementioned variables we controlled for the geographical location of the borrower to account for differences in the economic environment that might have affected the borrower. We grouped the borrowers’ states of residency into eight groups based on the Bureau of Economic Analysis classification, and added a special armed forces group for Military personnel serving overseas. Lastly, we included the final interest rate for each loan as a predictor in our model.⁸ Arguably, in a financially efficient world, this final interest rate, which was determined using a bidding process, should reflect all the information available to lenders

⁶ Because agreement across coders for age was lower, we also tested a model without this variable. Excluding the age variable did not qualitatively affect our results.

⁷ In addition to coding the demographics, we asked our judges to provide a score for attractiveness and trustworthiness for each borrower based on the picture (similarly to Pope and Sydnor 2011). However, given the high degree of disagreements across raters we decided not to use these measures in our analyses.

⁸ Because the maximum interest rate proposed by the borrower and the eventual lender’s rate are highly correlated we include only the lender’s rate in the model.

(including the textual information), thus our models test whether the text is predictive over and beyond that rate. However, we acknowledge that Prosper’s bidding mechanism may allow for some strategic behavior by sophisticated lenders, thus not fully reflecting a market efficient behavior (Chen, Ghose, and Lambert 2014). Table 1 presents summary statistics for the variables in our dataset.

*** Insert Table 1 about here ***

PREDICTING DEFAULT

Predictive Model (Stacking Ensemble)

Our objective in this section is to evaluate whether the text borrowers write in their loan request is predictive of their loan default up to three years post the loan request. In order to do so, we need to first build a strong benchmark—a powerful predictive model that includes the traditionally used financial and demographics information and maximizes the chances of predicting default using these variables. Second, we need to account for the fact that our model may include a very large number of predictors (over one thousand bi-grams). In evaluating a predictive model, it is common to compare alternative predictive models and choose the model that best predicts the desired outcome—loan repayment in our case. From a purely predictive point of view, a better approach, commonly used in machine learning, is to train several predictive models and rather than choosing the best model, create an ensemble or stack the different models. An ensemble of models benefits from the strength of each individual model and at the same time reduces the variance of the prediction. Accordingly, for the purpose of leveraging the textual information to predict default, we apply that approach.

The stacking ensemble algorithm includes two steps. In the first step, we train each model

on the calibration data. Because of the large number of textual variables in our model, we employ a simultaneous variable selection and model estimation in the first step. In the second step, we build a weighting model to optimally combine the models calibrated in the first step.

We consider four types of models in the first step. The models vary in terms of the classifier used and the approach to model variable selection. The four models include two logistic regressions and two versions of decision tree classifiers.⁹

Regularized Logistic Regressions (L1 and L2 Regularization)

The two logistic regressions are L1 and L2 regularization logistic regressions. These models differ with respect to the penalization terms for variable selection. The penalized logistic regression likelihood is:

$$L(Y|\beta, \lambda) = \sum_{t=1}^n (y_t \log(p(X_t|\beta)) + (1 - y_t) \log(1 - p(X_t|\beta))) - \lambda J(\theta),$$

where $Y = \{y_1, \dots, y_n\}$ is the set of binary outcome variables for n loans (loan repayment), $p(X_t|\beta)$ is the probability of repayment based on the logit model, where X_t is a vector of textual, financial and demographic predictors for loan t , β are a set of predictors' coefficients, λ is a tuning penalization parameter to be estimated using cross-validation on the calibration sample, and $J(\theta)$ is the penalization term. The L1 and L2 models differ with respect to the functional form of the penalization term, $J(\theta)$. In L1, $J(\theta) = \sum_{i=1}^k |\beta_i|$, while in L2, $J(\theta) = \sum_{i=1}^k \beta_i^2$, where k is the number of predictors. Whereas L1 tends to shrink many of the regression parameters to exactly zero and leave other parameters with no shrinkage, L2 tends to shrink many parameters to small but non zero values. Therefore, L1 is similar in spirit to the Lasso regression penalty and L2 to the ridge regression penalty. Before entering the variables into the L1 and L2 regression

⁹ We also considered a third type of decision tree (the AdaBoost tree method) but we dropped it due to poor performance on our data.

we standardize all variables (Tibshirani 1997).

Tree-based Methods (Random forest and Extra Trees)

The two tree-based methods we include in the ensemble are the random forest and the Extra Trees. The idea behind both models is to combine a large number of decision trees. Thus, to some extent, each of these tree-based methods is an ensemble in and of itself. In these models, trees are chosen to resolve misclassification of previously included trees. The random forest randomly draws with replacements subsets of the calibration data to fit each tree, and a random subset of features (variables) is used in each tree. The random forest approach mitigates the problem of over-fitting in traditional decision trees. The Extra Trees, on the other hand, is an extension of the random forest in which the thresholds for each variable in the tree are also chosen at random. Due to the size of the feature space, a χ^2 feature selection was first applied to the calibration data where the k-best features were kept. The optimal number of features to keep (k) was computed using an 80/20 split on the calibration data.

We used the scikit learn package in Python (<http://scikit-learn.org/>) to implement the four classifiers on a random sample of 80% of the calibration data. For the logistic regressions, we estimated the λ penalization parameter by grid search using a 3-fold cross validation on the calibration sample. For the tree-based methods, to limit over-fitting of the trees we randomized the parameter optimization (Bergstra and Bengio 2012) using a 3-fold cross validation on the calibration data to determine the structure of the tree (e.g., number leaves, number of splits, depth of the tree, and criteria). In the randomized parameter optimization, the parameters are sampled from a distribution (uniform) over all possible parameter values. We use a randomized parameter optimization rather an exhaustive search (or a grid search) due to the large number of variables in our model.

Model Stacking and Predictions

In the second step, we build a multinomial logit model to combine the ensemble of models using the remaining 20% of the calibration data. The probabilities from the second step multinomial logit model weigh the classification probabilities of our four approaches (the two logistic regularization regressions and the two decision trees methods) from the first step models, resulting in an overall weighted average of the different classifiers in the ensemble. We estimated an ensemble of the aforementioned four models, as well as subsets of these models and found that the ensemble with all the models performs the best. Accordingly, in the rest of this section we describe the predictions based on that ensemble.

To test our hypothesis that the text borrowers wrote in their loan requests is predictive of future default, we use a 10-fold cross validation. We randomly split the loans into 10 equally sized groups, calibrate the ensemble algorithm on nine groups and predict the remaining group. To evaluate statistical significance, we repeated the 10-fold cross validation 10 times, using different random seeds at each iteration. By cycling through the 10 groups and averaging the prediction results across the 10 cycles and 10 replications of the 10-fold cross validation we get a robust measure of prediction. Because there is no obvious cut-off for a probability from which one should consider the loan as defaulted, we use the “area under the curve” (AUC) of the Receiver Operating Characteristic (ROC) curve, a commonly used measure for prediction accuracy of binary outcomes. Additionally, because our interest is mainly in predicting defaults (as oppose to predicting loan repayment) we report the Jaccard index of loan default (e.g., Netzer et al. 2012; Toubia and Netzer 2016), which is defined as the number of correctly predicted defaulting loans divided by the loans that were defaulted but missed, loans that were predicted to be defaulted but were repaid and correctly predicted defaulted loans. This gives us an intuitive

measure of hit rates of defaulting loans penalized for erroneous predictions of both type I and type II errors. Finally, because base rate default varies significantly across credit grade levels, we report the predictions by splitting our sample to three groups of credit grades, high (AA, A), medium (B, C), and low (D, E, HR).

Empirical Results

We compare three versions of the stacking ensemble model: (1) an ensemble calibrated only on the financial and demographic data. This model mimics the loan default prediction models commonly used in the academic research and practice; (2) a model that includes just the textual information and ignores the financial and demographic information, and (3) a model that includes financial and demographic information together with the textual data. The comparison of models (2) and (3) informs us about the incremental predictive power of the textual information over and beyond the predictors commonly used in the financial industry. Comparing models (1) and (2) informs the degree of predictive information contained in the textual information relative to the financial and demographic information.

Table 2 details the average results of the 10-fold cross validation across 10 random shuffling of the observations. Figure 1 depicts the average ROC curve with and without the textual data for one representative 10-fold cross validation. The AUC is the area under the ROC curve, where a better predictive model is a model with a ROC curve that is closer to the upper left corner of the graph. The AUC of the model with both textual and financial and demographics information is 2.89% better than the AUC of the model with only financial information. This difference is statically significant as the model with both textual and financial and demographics information has higher AUC in all 100 replications of the cross validation exercise. The textual

information helps prediction for all credit grade levels, with the low and medium levels exhibiting the highest improvement, though at low credit grade level the improvement was not statistically significant (the model with textual information had higher AUC only in 83 out of the 100 replications of the cross validation exercise). Across credit levels we find that adding textual information improves default predictions over and beyond a model that is based only on financial and demographic information by 2.68-4.03%.

Interestingly, if we were to ignore the financial and demographic information and use only the borrower textual information, we obtain an AUC of 66.68% compared to an AUC of 70.52% for the model with only financial and demographic information. Thus, the textual information captures a large portion of the information commonly included in traditional measures.

*** Insert Table 2 and Figure 1 about here ***

We conducted a back-of-the-envelope calculation to quantify the managerial relevance and financial implications of the improvement in predictive ability offered by the textual data. Taking the approximately 275K loans granted by Prosper in 2015, and assuming an average default rate of 10% (the lower bound of most recently reported default rates), if a defaulter repays on average 25% of the loan before defaulting (based on estimates published by crowdfunding consulting agencies), the improvement in default prediction can lead to nearly \$3.2M funds a year that will not be allocated to defaulting borrowers. This amount can be further loaned and accumulate interest, thereby making it a conservative estimation. Saving that amount yearly is substantial in comparison to the total annual volume and the average size of loans at Prosper.

To summarize, the text borrowers write in their loan request can significantly help predict

loan default even when accounting for financial and demographic measures. In this section we employed an ensemble-based predictive model that aims to maximize predictive ability.

Unfortunately, these models provide little to no interpretation of their parameter estimates and the words and topics that predict default. Therefore, in the second part of this paper we present a series of analyses that shed light on the words and writing styles that were most likely to appear in defaulting loans.

WORDS, TOPICS, AND WRITING STYLES THAT ARE ASSOCIATED WITH DEFAULT

The result that text has a predictive ability similar in magnitude to the predictive ability of financial and demographic information is perhaps surprising. However, this result is consistent with the idea that people who differ in the way they think and feel also differ in what they say and write about those thoughts and feelings (Fast and Funder 2008; Hirsh and Peterson 2009; Schwartz et al. 2013; Yarkoni 2010a). We employed four approaches to uncover whether words, topics, and writing styles of defaulters differ from those who repaid their loan. First, we use a naïve Bayes classifier to identify the combination of words that most distinguish defaulted from fully-paid loans. The advantage of the naïve Bayes is in providing intuitive interpretation of the words that are most discriminative between defaulted and repaid loans; however, its disadvantage is that it assumes independence across predictors and therefore cannot control for the financial and demographics variables (or for the dependence among the textual variables). To alleviate this concern, we use a logistic regression with L1 penalization to uncover the words that are more associated with default after controlling for the financial and demographic information. Results of the L1 regression are qualitatively similar to those of the naïve Bayes approach, and therefore we discuss the naïve Bayes here and the L1 regression in the Web Appendix. The

convergence of results across the naïve Bayes and L1 regression increases our confidence in these findings. To look beyond specific words or bi-grams and into the topics discussed in each loan and writing styles employed, we use a latent Dirichlet allocation analysis. Based on this analysis we uncovered three type of topics, those related to the loan purpose (e.g., business loan or a collage loan), those related to the borrower circumstances, and those related to pleading to lenders. Finally, we employ a well-known dictionary, the Linguistic Inquiry and Word Count (LIWC; Tausczik and Pennebaker 2010), to identify the writing styles that are most correlated with defaulting or repaying the loan.

Words that Distinguish between Loan Requests of Paying and Defaulting Borrowers

To investigate which words in the loan application most discriminate between borrowers who default and borrowers who repay the loan in full, we ran a naïve Bayes classifier using the Python NLTK 3.0 package. The naïve Bayes classifier uses Bayes rule and the assumption of independence among words to estimate each word’s likelihood of appearing in defaulted and non-defaulted loans. As with the ensemble approach we ran the naïve Bayes classifier on bi-grams (all possible words and pairs of words) that appeared in at least 400 loans (1,032 bi-grams). We then calculate the most “informative” bi-grams in terms of discriminating between defaulted and non-defaulted loans by calculating the bi-grams with the highest ratio of $P(\text{bi-gram}|\text{defaulted})/P(\text{bi-gram}|\text{repaid})$ and the highest ratio of $P(\text{bi-gram}|\text{repaid})/P(\text{bi-gram}|\text{defaulted})$.

Table A2 in the Web Appendix presents the lists of words and their likelihood of appearing in defaulted versus paid loan requests for words that had ratios larger than 1:1.1. To better visualize the results Figures 2 and 3 present word clouds of the naïve Bayes analysis of bi-

grams in their stemmed form. The size of each bigram in Figures 2 and 3 corresponds to the likelihood that the bigram will be included in defaulted loan request versus a repaid loan request in Figure 2 and in a repaid loan request versus defaulted loan request in Figure 3. For example, the bi-gram “all_bill” in Figure 2 is 2.9 times more likely to appear in a paid-back than a defaulted loan request, while the word “god” is 2.2 times more likely to appear in defaulted than paid-back loan requests. The central cloud in each figure presents the most discriminant bi-grams (cutoff=1.5 for Figure 2 and cutoff=1.6 for Figure 3) and the satellite clouds represent emerging themes based on our interpretation of groups of words that had high discriminant value.

*** Insert Figures 2, 3 around here ***

Several insights can be gained from this analysis. Relative to defaulters, borrowers who paid in full were more likely to include in their loan application (i) Words associated with their financial situation such as “all_bill,” “card_with,” and “car_insurance.” (ii) Words that may be a sign of projected improvement in financial ability: “promotion,” “graduating,” and “wedding.” (iii) Relative words such as “than_the,” “rather,” and “more_than.” (iv) Time related words such as “year_now,” “three_year,” and “annual.” The above indicates that borrowers who paid in full may have nothing to hide, a promising financial future, and a seemingly complex story (as indicated by the relative words), which past research suggest is likely to be truthful. The use of relative and time words has been associated with greater candor because honest stories are usually more complex (Newman et al. 2003). Dishonest stories, on the other hand, are simpler, allowing the lying storyteller to conserve cognitive resources in order to focus on the lie more easily (Tausczik and Pennebaker 2010). Some words in Figure 2 are idiosyncratic to Prosper (such as “reinvest” which means these people would like to borrow funds in order to lend them to others on Prosper), and hence are not discussed.

Turning to Figure 3, not surprisingly, borrowers who defaulted were more likely to mention words related to (i) Financial hardships (“payday_loan,” “child_support,” and “bankruptcy”) and general hardship (“divorce,” “emergency,” and “stress.”) This result is in line with Herzenstein, Sonenshein, and Dholakia (2011) who found that discussing personal hardship in the loan application is associated with borrowers who are late on their loan payments. (ii) Explaining their situation (“loan_explain,” “explain_what,”) and discussing their work state (“been_work,” “work_hard”). Providing explanations is often connected to past deviant behavior (Michels 2011; Sonenshein, Herzenstein, and Dholakia 2011). (iii) Appreciative and good-manner words toward lenders (“God_bless,” “and_thank”) and pleading lenders for help (“need_help,” “help_get.”) (iv) Referring to external sources such as “God,” “daughter,” or “husband.” The strong reference to others has been shown to exist in deceptive language style. Liars tend to avoid mentioning themselves, perhaps to distance themselves from the lie (Hancock et al. 2007; Newman et al. 2003). With respect to the frequent mention of God in defaulting loans, Kupor, Laurin, and Levav (2015), find that reminders of God can increase the likelihood of people engaging in riskier behaviors, alluding to the possibility these borrowers took a loan they were unable to repay. (v) Time related words (“few_month,” “month_that”) and future tense words (“would_use,” “payment_will.”) While both paying and defaulting borrowers use time related words, defaulters seem to focus on the shorter term (a month,) while repayers on the longer term (a year). This result is consistent with the finding of Lynch et al. (2010), who showed that long-term planning (as opposed to short planning) was associated with lower procrastination and higher degree of assignment completion. Further, the degree of long-term planning was associated with FICO scores. The mention of shorter horizon time words by defaulters is also consistent with Shah, Mullainathan, and Shafir (2012) finding that financial

resource scarcity leads people to shift their attention to the near future, neglecting the distant future, hence leading to over-borrowing. The above words suggest that defaulting borrowers attempted to garner empathy from lenders, seem forthcoming and appreciative, but when the time to repay the loan came they were unable to escape their reality.

In order to investigate whether the results of the naïve Bayes analysis are sensitive to the inclusion of demographics and financial information and the interdependence among words we employ a logit regression with an L1 penalization with same 1,032 bi-grams used in the ensemble learning and naïve Bayes analysis as well as the demographic and financial information. This analysis, while less easily interpretable than the naïve Bayes, provided very similar results, and is reported in Web Appendix (see Tables A3 and A4). The L1 regression results confirm that the writing styles and intentions we identified through the naïve Bayes analysis are not merely a proxy of the demographic and financial information.

Analyzing the Topics Discussed in Each Loan Request and Their Relationship to Default

The previous analyses allowed meaningful insights into the discriminative power of specific words or bi-grams between repaid and defaulted loans. In Figures 2 and 3 we grouped the individual words into topics based on our interpretation and judgment. However, this method cannot provide insights into the type of topics that were discussed in loan requests. Hence we employed a latent Dirichlet allocation analysis (LDA; Blei, Ng, and Jordan 2003).

LDA is the most commonly used mixed membership model. LDA assumes that each document (loan request in our case) is constructed of a mixture of multiple topics (with a Dirichlet priors). The words in the document are probabilistically related to each one of the topics. For example, the words “university,” “graduate,” and “school” were associated with high

probability with one topic: college. Similar to factor analysis, the same word can have high likelihood of appearing in multiple topics. LDA assumes that the researcher knows a-priori the number of topics-groups of associated words. These topics are collections of words that co-occurs with each other to create coherent groups. The LDA inference procedure tries to find these co-occurring words by identifying groups of words that appear together over and beyond chance across documents. As part of the inference procedure LDA also calculates the topic proportion—the percentage of each topic-group of associated words in each textual loan description. The sum of all these topics proportions in any given document equals to one.

We use the online variational inference algorithm for the LDA training (Hoffman, Bach and Blei 2010), following the settings and priors described in Giffith and Steyvers (2004). We used the 5,000 word stems that appeared most frequently across loan requests, eliminating infrequent words mitigates the risk of rare-words occurrences and co-occurrence confounding the topics. Because the LDA analysis requires the researcher to determine the number of topics to be analyzed, we varied the number of topics between two and 30, and used model fit, predictive ability, and interpretation of the topics to determine the final number of topics. We split the data into 80% calibration and 20% validation, and used the validation sample AUC of predicting loan default for a model with the financial and demographic information as well as the LDA topics varying from 2-30 topics. As can be seen in Figure 4, the model with 13 topics provided a good balance between predictive ability and model complexity. At 13 topics the improvement in predictive ability tapers off. To evaluate the LDA model fit, we use the commonly used perplexity measure. Similar to model fit measures such AIC or BIC, lower perplexity implies better fit. Based on perplexity alone one would choose a model with only four topics (perplexity = 74.6). However, the perplexity of the LDA model with 13 topics, which was

chosen based on its interpretability and ability to predict default, is only slightly higher (perplexity=75.4). Finally, from an interpretation point of view the model with 13 topics leads to easily interpretable topics. Similar approach of balancing topic interpretation and model fit has been taken in previous LDA studies (e.g., Chang et al. 2009; Hansen, McMahon, and Prat 2014).

The 13 topics we identify can be roughly divided into three categories, the main reason for requesting the loan (e.g., school loan, mortgage, business loan), the borrower life circumstances (e.g., the borrower financial or medical condition), and how borrowers plead to lenders in the loan application (e.g., words related politeness or providing explanations). We note that some topics can be interpreted as both loan purpose and life circumstances. For example, a borrower may mention school because this is the purpose of the loan, or because she wants to highlight her education to improve her chance of obtaining the loan or obtaining better rate. Table 3 presents the 13 topics and the words that are most representative of the topic based on the relevance score (Sievert and Shirley 2014), and Table A5 in the Web Appendix includes a more comprehensive list of the top 30 words with the highest relevance measure for each topic. The relevance score for topic k and word w is calculated as:

$$r(k, w) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log \left(\frac{\phi_{kw}}{P_w} \right),$$

where, λ is the weight given to topic k under word w relative to its lift, ϕ_{kw} is probability of word w appearing in topic k , and P_w is the frequency in which word w appears across documents. Thus, the relevance measure balances the prominence of a word in a topic to its prominence in the entire corpus. For the purpose of topic interpretation, we use $\lambda = 0.5$.

*** Insert Table 3 around here ***

We find four loan purpose topics: relocation loans, school loans, rate reduction loans, and business loans. Additionally, we find seven life circumstances topics: family medical issues

(could also be a loan purpose), monthly income details, housing descriptions, details about one's credit score, details of prior loans on Prosper, and two different topics highlighting monthly expenses. The monthly expenses topics are most likely related to a set of expenses that Prosper recommended borrowers mention as part of their loan request during our data period. Finally, we find two topics related to how borrowers address lenders: pleading for help/asking for a second chance, and providing explanations.

To relate the topics mentioned in each loan request's text to the likelihood of default, we ran a binary logit regression with loan repayment = 1 and default = 0 as the dependent variable, the probability of each topic appearing in the loan based on our LDA analysis (the topic Monthly Expenses 1 serves as benchmark), and the same set of textual information metrics as well as the financial and demographic variables used in the ensemble learning and L1 regularization logistic regression described earlier. Table 4, presents the results of the binary logit regression with the LDA topics. First, looking at the financial variables, we find that all parameter estimates are significant and in the expected direction. That is, repayment likelihood is increasing as credit grades improve, but decreasing with debt to income ratio and home ownership. As expected repayment likelihood is negatively correlated with higher lender rate, suggesting some level of efficiency among Prosper lenders. Finally, higher dollar amount loans were more likely to be defaulted.

*** Insert Table 4 around here ***

More relevant to the current analysis, is the relationship between the topics identified and loan repayment. Relative to the topic Monthly Expenses 1, we find that topics of Relocation Loans, Collage Loans, and Rate Reduction loans all have higher likelihood of repayment. Business Loans on the other hand have lower likelihood of repayment. This could be due to the slow economy and recession that occurred during the time borrowers in our sample had to pay

back the loan (2008-2010). Consistent with the naïve Bayes analysis we find that pleading for help explaining one's financial and life circumstances such as medical conditions and explanations are associated with lower repayment likelihood. We also find that tendency to detail the monthly expenses and financials is associated with higher likelihood of repayment, perhaps because providing such information is indeed truthful and forthcoming (having nothing to hide).

Although the purpose of the LDA analysis was to learn about the topics discussed in loan requests rather than to predict default, we nevertheless tested the predictive ability of the uncovered topics. We find that the model that includes the LDA topics fits the data better than a model that does not include the textual information in terms of the Akaike information criterion ($AIC_{LDA} = 20,819$ and $AIC_{notext} = 21,078$). Furthermore, the likelihood ratio test significantly supports the model with the textual information relative to the model without the textual information ($LR_{DF=17} = 292.78, p < 0.001$). We ran a 10-fold cross validation similar to the one conducted for the ensemble learning model. We find that the model with the LDA topics and the other textual variables (e.g., number of characters in the loan request) predicts defaults better than a baseline model that includes all the financial and demographic information but no textual information ($AUC_{LDA} = 71.8\%$ vs. $AUC_{noLDA} = 70.1\%$). The model with the LDA variables provided higher AUC relative to the model without the textual information in all 10 folds. In sum, using multiple methods we uncovered themes and words that differentiate defaulted from fully paid loans. Next, we use a well-researched dictionary to explore borrowers' *writing styles* with the objective of *exploring* the personality and emotional state of those who defaulted.

Intentions, Circumstances, and Personalities of Those Who Defaulted

In this section we rely on one of the more researched and established text analysis

methods, the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker, Booth, and Francis 2007; Tausczik and Pennebaker 2010). This dictionary groups almost 4,500 words into 64 linguistic and psychologically meaningful categories such as tenses (past, present, future), forms (I, we, you, she or he), social, positive, and negative emotions. Since its release in 2001, many researchers have examined and employed it in their research (see Tausczik and Pennebaker (2010) for a comprehensive overview). For example, word usage has been associated with the text writer's personality, focusing especially on the big five personality traits—extraversion, agreeableness, conscientiousness, neuroticism, and openness (Beukeboom, Tanis, and Vermeulen 2013; Kosinski, Stillwell, and Graepel 2013; Pennebaker and Graybeal 2001; Pennebaker and King 1999; Yarkoni 2010a; Schwartz et al. 2013), physical and mental health (Pennebaker 1993; Preotiuc-Pietro et al. 2015), age and gender (Pennebaker and Stone 2003; Schwartz et al. 2013), emotional state (Pennebaker, Mayne, and Francis 1997), and deception (Newman et al 2003).

We note that other dictionaries are common in the interpretation of financial related text (e.g., quarterly and annual companies' reports), such as DICTION and the Loughran and McDonald (2011)'s list of negative and positive words. However, because these dictionaries were built around professional financial writing and are relevant to financial conditions of companies, they are less appropriate for our data, which is written by individuals, with often little financial background, and is aimed to provide information about individuals rather than companies. Further, the list of 2,000 topics developed by Schwartz et al. (2013) is also inadequate for our purposes because it is based on Facebook (and later used successfully on Twitter; Preotiuc-Pietro et al. 2015) and thus includes many items related to parting, drinking, and swearing—words that rarely appear in the text we analyze.

Because the same word may appear in several dictionaries, we first calculated the proportion of stemmed words in each loan request that belong to each of the 64 dictionaries.¹⁰ We then estimated a binary logit model to relate the proportions of words in each loan that appear in each dictionary to whether the loan was repaid (as opposed to defaulted; load repaid=1 and loan defaulted=0), controlling for all financial and demographic variables used in the stacking ensemble, the L1 regularization logistic regression model, and the LDA analysis described earlier. The estimates of the binary logit model with LIWC dictionaries as covariates are presented in Table 5.

*** Insert Table 5 around here ***

The objective of this analysis is to infer which aspects of the borrower's writing style are significantly correlated with loan repayment/default after controlling for the financial and demographic information. First, we note that all of the financial and demographic control variables are in the expected direction and consistent with those of the LDA analysis described earlier.

Fourteen of the sixty-four LIWC dictionaries were significantly related to repayment behavior. Several of them corroborate our previous results from the naïve Bayes and L1 analyses. More importantly, relating our findings to previous research that leveraged the LIWC dictionary, we observe that defaulted loan requests contain words that are associated with the writing style of liars and of extroverts.

We begin with deception. Looking at Table 5, we see that the following LIWC sub-dictionaries, that have been shown to be associated with greater likelihood of deception, are associated in our analysis with greater likelihood to default: (1) present and future tense words.

¹⁰ For this analysis we did not remove words with less than three characters and infrequent words as we are matching words to pre-defined dictionaries.

This result is similar to our findings from the naïve Bayes and past research (Pasupathi 2007); (2) motion words (e.g., “drive,” “go,” and “run.”) Newman et al. (2003) show that increased usage of motion words is indicative of lying because deceptive communications are less cognitively complex. Indeed, the words in this dictionary are simple verbs that are associated with less cognitive complexity. Conversely, relative words (e.g., “closer,” “higher,” and “older”) have been associated with greater truthfulness. Relative words are used to make comparisons, resulting in more complex stories (Pennebaker and King 1999). As expected, and consistent with the results from the naïve Bayes analysis, we find that greater use of these words is associated with higher likelihood of paying back the loan; (3) Similar to our finding from the naïve Bayes analysis that defaulters tend to refer to others, we find that social words (e.g., “mother,” “father,” “he,” “she,” “we,” and “they”) are associated with higher likelihood of default. Along these lines, Hancock et al. (2007) showed that linguistic writing style of liars is reflected by lower use of first person singular and higher use of first person plural such as “we” (See also Newman et al. 2003). We note, though, that in the context of hotel reviews, Ott, Cardie, and Hancock (2012) find higher use of “I” in fake reviews possibly to increase reliability, perhaps alluding to the dissemination of the aforementioned results among professional liars; (4) time words (e.g., “January,” “Sunday,” “morning,” and “never”) and space words (e.g., “above,” “inch,” and “north”) were associated with higher likelihood of default. These words have been found to be prevalent in deceptive statements written by prisoners (Bond and Lee 2005);

Taken together, we find that several of the LIWC dictionaries that have been previously found to associate with deception are also negatively associated with loan repayment (positively associated with loan default). We wish to note that we do not claim that borrowers employing these word-themes are outright lying. Rather, we believe that, consciously or not, the text they

wrote has traces of dishonesty, or a sort of online “involuntary sweat”.

Our second observation is that the sub-dictionaries associated with the writing style of extroverts are also associated with greater likelihood to default. The premise that the text may be indicative of deeper traits and emotional states is predicated on the idea that there is a systematic relationship between the words people use and individuals’ personality traits (Fast and Funder 2008; Hirsh and Peterson 2009; Yarkoni 2010b), identities (McAdams 2001), and emotional states (Tausczik and Pennebaker 2010). The relationship between word usage and personality traits has been found across multiple textual media such as essays about the self (Hirsh and Peterson 2009), personal blogs (Yarkoni 2010a), social media (Schwartz et al. 2013), and naturalistic recordings of daily speeches (Mehl, Gosling, and Pennebaker 2006). The reasons for this relationship stems from the human tendency to tell stories and express internal thoughts and emotions through these stories, which are essentially made possible by language.

Extroverts have been shown to use more religious and body related words (e.g., “mouth,” “rib,” “sweat,” and “naked”; Yarkoni 2010a), social and humans words (e.g., “adults,” “boy,” and “female”; Hirsh and Peterson 2009; Pennebaker and King 1999; Schwartz et al. 2013; Yarkoni 2010a), and motion words (e.g., “drive,” “go,” and “run”; Schwartz et al. 2013)—all of which are significantly related to a greater likelihood of default in our analysis (see Table 5). Further, extroverts use more achievement word (e.g., “able,” “accomplish,” and “master”) and less filler words (e.g., “blah” and “like”; Mairesse et al. 2007). Consistent with these findings, we find that achievement words are significantly and positively associated with default while filler words are positively associated with repayment. The finding that defaulters are more likely to exhibit writing style of extroverts is consistent with research showing that extroverts are more likely to take risks (Nicholson et al. 2005), engage in compulsive buying of lottery tickets

(Balabanis 2002), and are less likely to save (Brandstätter 2005; Nyhus and Webley 2001).

It is not a coincidence that the fourteen LIWC dictionaries that were significantly correlated with default are also correlated with extroversion and/or deception. Past literature has consistently documented that extroverts are more likely to lie, and not only because they talk to more people but rather because these lies help smooth their interactions with others (Weiss and Feldman 2006).

While the LIWC dictionaries have been shown to be correlated with the other big five personality traits, we did not find consistent and conclusive relationship between the dictionaries associated with each of the other personality traits and loan repayment. Similarly, results from other research on the relationship between LIWC and gender, age, mental and emotional states did not consistently relate to default in our study.

In summary, relying on previous research on LIWC, we find that loan requests written in a manner consistent with the writing style commonly found in writings of extroverts and liars, are more likely to eventually default, sometime years after the loan request was crafted. We acknowledge that there may be variables that are confounded with both the observable text and unobservable personality traits or states we discussed that are accountable for the repayment behavior. Nevertheless, from a predictive point of view, we find that the model that includes the LIWC dictionaries fits the data better than a model that does not include the textual information in terms of the Akaike information criterion ($AIC_{\text{text}} = 20,900$ and $AIC_{\text{notext}} = 21,078$). Furthermore, the likelihood ratio test significantly supports the model with the textual information relative to the model without the textual information ($LR_{DF=69} = 319.94$, $p < 0.001$). To test for the predictive ability of this model we ran a 10-fold cross validation similar to the one conducted for the ensemble learning model. We find that the model with LIWC predicts defaults

better than a baseline model that includes all the financial and demographic information but no textual information ($AUC_{LIWC} = 70.9\%$ vs. $AUC_{noLIWC} = 70.1\%$). The model with the LIWC variables provided higher AUC than the model without the textual information in all 10 folds.

GENERAL DISCUSSION

The words we write matter. Aggregated text has been shown to predict market trends (Bollen, Mao, and Zeng 2011) and behaviors of individual stocks (Tirunillai and Tellis 2012), market structure (Netzer et al. 2012), virility of news articles (Berger and Milkman 2012), prices of the discussed services (Jurafsky et al. 2014), and political elections (Tumasjan et al. 2010). At the individual text writer level, text has been used to evaluate the state of mind of email writers (Ventrella 2011), to identify liars (Newman et al. 2003) and fake hospitality reviews (Ott, Cardie, and Hancock 2012), to assess the personality traits of bloggers (Yarkoni 2010a) and Facebook users (Schwartz et al. 2013), and the mental state of those who Tweet (Preotiuc-Pietro et al. 2015). In this paper, we show that text has the ability to predict financial behavior of its writer in the distant future with significant accuracy.

Using data from an online crowdfunding platform we show that incorporating the text borrowers write in their loan application into traditional models that predict loan default based on financial and demographic information about the borrower significantly and substantially increases the ability of these models to predict default. We then analyzed using naïve Bayes analysis, L1 regularized regression, and LDA analysis, the words and topics borrowers included in their loan request and found that at the time of loan application, defaulters used simple but wordier language, wrote about hardship, further explained their situation and why they need the loan, and tended to refer to other sources such as their family, god, and chance. Building on past

research and commonly used LIWC dictionary we were able to infer that defaulting borrowers write similarly to people who are extroverts and to those who lie. Importantly, these results were obtained after controlling for the borrower's credit grade, which should capture the financial implications of the borrower's life circumstances, and the interest rate given to the borrower, which should capture difference in the risk of different types of loans.

Theoretical and Practical Contribution

Our research makes the following theoretical and practical contributions. First, in an environment characterized by high uncertainty, we find that verifiable and unverifiable data have similar predictive ability. While borrowers can truly write whatever they wish in the textbox of the loan application—supposedly “cheap talk” (Farrell and Rabin 1996)—their word usage is predictive of future repayment behavior at a similar scale as their financial and demographic information. This finding implies that whether it is intentional and conscious or not, borrowers' writings seem to disclose their true nature, intentions, and circumstances. This finding contributes to the literature on implication and meaning of word usage (Fast and Funder 2008, Hirsh and Peterson 2009; Preoțiu-Pietro et al. 2015; Schwartz et al. 2013; Yarkoni 2010a) by showing that people with different economic and financial situations use words differently.

Second, we contribute to the text analytics literature by showing that word usage and writing styles are predictive of future behaviors of the text writer, months and even years after the text was written. The automatic and manual text-mining literature has primarily concentrated on predicting behaviors that occur at the time of writing the text, such as lying about past events (Newman et al. 2003) or writing a fake review (Ott, Cardie and Hancock. 2012), but not on predicting future behavior of writers. In one interesting exception, Slatcher and Pennebaker

(2006) show that couples who used more positive emotion words when texting to each other were shown to be more likely to continue dating three months later. Our approach to predicting default relies on an automatic algorithm that mines individual words (including those without much meaning such as articles and fillers) in the entire set of textual corpora. There have been some work distilling from text the narrative used to facilitate economic transaction (e.g., Chen, Yao, and Kotha 2009; Herzenstein, Sonenshein, and Dholokia 2011; Martens, Jennings, and Jennings 2007), but these approaches are laborious (human readings of the textual information) and not scalable, which limit their predictive ability and practical use.

Third, we provide evidence that our method of automatically analyzing free text is an effective way of replacing some aspects of the human interaction of traditional bank loans. Furthermore, because lending institutions place a great deal of emphasis on the importance of models for credit risk measurement and management, they have historically developed their own proprietary models, which often have a high price tag due to data acquisition (Bloomberg 02/2013). Collecting text may be an effective and low cost supplement to the traditional financial data and default models. That being said, although our objective in this research is to explore the predictive and informative value of the text in loan applications, using such information to decide to whom organizations should grant loans may carry ethical and legal considerations within the restrictions of this highly regulated industry. For example, using textual information to discriminate among group members may violate the fairness in classification (Dwork et al. 2012), and to the extent that the textual information is directly related to race, color, religion, national origin, gender, marital status, or age, the use of such information may be considered unlawful based on the 1974, Equal Credit Opportunity Act. Institutional judgement regarding the usage of our findings is beyond the scope of our research.

Avenues for Future Research

Our research takes the first step in automatically analyzing text in order to predict default, and therefore initiates multiple research opportunities. First, we focus on predicting default because this is an interesting behavior that is less idiosyncratic to the crowdfunding platform whose data we analyze (compared with lending decisions). Theoretically, many aspects of loan repayment behavior, which are grounded in human behavior (e.g., extroversion (Nyhus and Webley 2001) or demographics (Karlan 2005)), should be invariant to the type of loan, and the platform used for the loan, whereas other aspects may vary from one context to another. The robustness of our results should be tested with different populations, other types of unsecured loans, such as credit card debt, as well as secured loans, such as mortgages.

Second, we focus on predicting loan default because we believe it is a behavior of high financial relevance, in which the text borrower write can have subtle traces that can be helpful in predicting current circumstances and intentions as well as future behavior. We encourage future research to explore the effect of the text in the loan request also on the lenders' behavior in terms of granting the loan and bidding a particular interest rate. Some researchers have studied these questions, but on a much smaller scale compared with our endeavor (e.g., Herzenstein, Dholakia, and Sonenshein 2011; Michels 2012).

Third, our results should be validated in and extended to other types of media, such as phone calls or online chats. It would be interesting to test how an active conversation—two-sided correspondence—versus only one-sided input as in our data (only borrowers write, lenders are not involved at the time of the loan application) may affect the results. One big difference is that in our data the content of the text is entirely up to the borrower—he or she discloses what they

wish in whichever writing style they choose. In a conversation, the borrower may be prompted to provide certain information.

Forth, we document specific bi-grams and themes that might help lenders avoid defaulting borrowers, and help borrowers better express themselves in requesting the loan. Based on the market efficiency hypothesis, if both lenders and borrowers internalize the results we documented, these results may change. However, evidence from body language research and deception detection mechanisms suggest that such effects rarely fully disappear.

Finally, while we are studying the predictive ability of written text regarding a particular future behavior (loan default), our approach can be easily extended to other behaviors and industries. For example, universities and business schools might be able to predict students' success based on the text in the application (beyond human manual reading the essays). Similarly, human resource practitioners and recruiters can use the words in the text applicants write to identify promising candidates.

To sum, we find that borrowers leave meaningful traces in the text of loan applications that help predict default, sometimes years post the loan request. We see this research as a first step in utilizing text mining to better understand and predict financial decision-making and outcomes, and consumer behavior more generally.

References

- Agarwal, Sumit, and Robert Hauswald (2010), "Distance and Private Information in Lending," *Review of Financial Studies*, 23 (7), 2757-2788.
- Avery, Robert B., Paul S. Calem, and Glenn B. Canner (2004), "Consumer Credit Scoring: Do Situational Circumstances Matter?," *Journal of Banking and Finance*, 28 (4), 835-856.
- Balabanis, George (2002), "The Relationship between Lottery Ticket and Scratch-Card Buying Behaviour, Personality and Other Compulsive Behaviours," *Journal of Consumer Behaviour*, 2 (1), 7-22.
- Bergstra, James, and Yoshua Bengio (2012), "Random Search for Hyper-Parameter Optimization." *Journal of Machine Learning Research*, 13 (Feb), 281-305.
- Berger Jonah, and Katherine L. Milkman (2012), "What Makes Online Content Viral?" *Journal of Marketing Research*: 49 (2), 192-205.
- Berger, Sven C., and Fabian Gleisner (2009) "Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending." *BuR-Business Research 2.1*, 39-65.
- Beukeboom, Camiel J., Martin Tanis, and Ivar E. Vermeulen (2013), "The Language of Extraversion Extraverted People Talk More Abstractly, Introverts are More Concrete," *Journal of Language and Social Psychology*, 32 (2), 191-201.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan), 993-1022.
- Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011), "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2 (1), 1-8.
- Bond, Gary D., and Adrienne Y. Lee (2005), "Language of Lies in Prison: Linguistic Classification of Prisoners' Truthful and Deceptive Natural Language," *Applied Cognitive Psychology*, 19 (3), 313-329.
- Brandstätter, Hermann (2005), "The Personality Roots of Saving—Uncovered from German and Dutch Surveys," In *Consumers, Policy and the Environment A Tribute to Folke Ölander*, 65-87, Springer.
- Chen, Ning, Arpita Ghosh, and Nicolas S. Lambert (2014) "Auctions for Social Lending: A Theoretical Analysis," *Games and Economic Behavior*, 86, 367-391.

- Chen, Xiao-Ping, Xin Yao, and Suresh Kotha (2009), "Entrepreneur Passion and Preparedness in Business Plan Presentations: A Persuasion Analysis of Venture Capitalists' Funding Decisions," *Academy of Management Journal*, 52 (1), 199-214.
- Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei (2009), "Reading Tea Leaves: How Humans Interpret Topic Models," *In Advances in Neural Information Processing Systems*, 288-296.
- DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper (2003), "Cues to Deception," *Psychological Bulletin*, 129 (1), 74-118.
- Dogra, Keshav and Olga Gorbachev (2016), "Consumption Volatility, Liquidity Constraints and Household Welfare," *The Economic Journal*, forthcoming.
- Duarte, Jefferson, Stephan Siegel, and Lance Young (2012), "Trust and Credit: The Role of Appearance in Peer-To-Peer Lending," *Review of Financial Studies*, 25 (8), 2455-2484.
- Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel (2012), "Fairness through Awareness." *In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214-226. ACM.
- Fast, Lisa A., and David C. Funder (2008), "Personality as Manifest in Word Use: Correlations with Self-Report, Acquaintance Report, and Behavior," *Journal of Personality and Social Psychology*, 94 (2), 334-346.
- Farrell, Joseph, and Matthew Rabin (1996), "Cheap Talk," *The Journal of Economic Perspectives*, 10 (3), 103-118.
- Gross, Jacob P.K., Cekic Osman, Don Hossler, and Nick Hillman (2009), "What Matters in Student Loan Default: A Review of the Research Literature." *Journal of Student Financial Aid*, 39 (1), 19-29.
- Hancock, Jeffrey T., Lauren E. Curry, Saurabh Goorha, and Michael Woodworth (2007), "On Lying and Being Lied to: A Linguistic Analysis of Deception in Computer-Mediated Communication," *Discourse Processes*, 45 (1), 1-23.
- Hansen, Stephen, Michael McMahon, and Andrea Prat (2014), "Transparency and Deliberation within the FOMC: a Computational Linguistics Approach." (2014), Working Paper, Columbia University.
- Harkness, Sarah K. (2016), "Discrimination in Lending Markets: Status and the Intersections of Gender and Race," *Social Psychology Quarterly*, 79 (1), 81-93.

- Herzenstein, Michal, Utpal M. Dholakia, and Rick L. Andrews (2011), "Strategic Herding Behavior in Peer-to-Peer Loan Auctions," *Journal of Interactive Marketing*, 25 (1), 27-36.
- Herzenstein, Michal, Rick L. Andrews, Utpal M. Dholakia, Evgeny Lyandres (2008), "The Democratization of Personal Consumer Loans? Determinants of Success in Online Peer-to-Peer Lending Communities," *Working Paper*, University of Delaware.
- Herzenstein, Michal. Scott Sonenshein, and Utpal M. Dholakia (2011), "Tell Me a Good Story and I May Lend You My Money: The Role of Narratives in Peer-To-Peer Lending Decisions," *Journal of Marketing Research Special Issue on Consumer Financial Decision Making*, 48, S138-S149.
- Hirsh, Jacob B., and Jordan B. Peterson (2009), "Personality and Language Use in Self-Narratives," *Journal of Research in Personality*, 43 (3), 524-527.
- Hoffman, Matthew, Francis R. Bach, and David M. Blei (2010) "Online Learning for Latent Dirichlet Allocation," *Advances in Neural Information Processing Systems*, 856-864.
- Griffiths, Thomas L., and Mark Steyvers (2004) "Finding Scientific Topics," *Proceedings of the National Academy of Sciences*, 101(1), 5228-5235.
- Iyer, Rajkamal, Asim Ijaz Khwaja, Erzo FP Luttmer, and Kelly Shue (2016), "Screening Peers Softly: Inferring the Quality of Small Borrowers," *Management Science*, forthcoming.
- Jurafsky, Dan, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith (2014), "Narrative Framing of Consumer Sentiment in Online Restaurant Reviews," *First Monday*, 19 (4).
- Karlan, Dean S. (2005), "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions," *American Economic Review*, 95 (5), 1688-1699.
- Kosinski, Michal, David Stillwell, and Thore Graepel (2013), "Private Traits and Attributes are Predictable from Digital Records of Human Behavior," *Proceedings of the National Academy of Sciences*, 110 (15), 5802-5805.
- Kupor, Daniella M., Kristin Laurin, and Jonathan Levav (2015), "Anticipating Divine Protection? Reminders of God Can Increase Nonmoral Risk Taking," *Psychological Science*, 26 (4), 374-384.
- Lewis, Michael (2015), *The Big Short: Inside the Doomsday Machine*. WW Norton & Company.
- Loughran, Tim, and Bill McDonald (2011), "When is a Liability Not a Liability? Textual

- Analysis, Dictionaries, and 10-Ks,” *The Journal of Finance*, 66 (1), 35-65.
- Luo, Chunyu, Hui Xiong, Wenjun Zhou, Yanhong Guo, and Guishi Deng (2011), “Enhancing Investment Decisions in P2P Lending: An Investor Composition Perspective,” In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 292-300.
- Lynch, John G., Richard G. Netemeyer, Stephen A. Spiller, and Alessandra Zammit (2010), "A Generalizable Scale of Propensity to Plan: the Long and the Short of Planning for Time and for Money." *Journal of Consumer Research* 37(1), 108-128.
- Mairesse, François, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore (2007), “Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text,” *Journal of Artificial Intelligence Research*, 30, 457-500.
- Martens, Martin L., Jennifer E. Jennings, and P. Devereaux Jennings (2007), “Do the Stories They Tell Get Them the Money They Need? The Role of Entrepreneurial Narratives in Resource Acquisition,” *Academy of Management Journal*, 50 (5), 1107-1132.
- Massolution CL (2015), “Crowdfunding Industry Report”
- Mayer, Christopher, Karen Pence, and Shane M. Sherlund (2009), “The Rise in Mortgage Defaults." *The Journal of Economic Perspectives*, 23 (1), 27-50.
- McAdams, Dan P. (2001), “The Psychology of Life Stories,” *Review of General Psychology*, 5(2), 100-123.
- Mc Laughlin, G. Harry (1969), “SMOG Grading—A New Readability Formula,” *Journal of Reading*, 12 (8), 639-646.
- Mehl, Matthias R., Samuel D. Gosling, and James W. Pennebaker (2006), “Personality in its Natural Habitat: Manifestations and Implicit Folk Theories of Personality in Daily Life,” *Journal of Personality and Social Psychology*, 90 (5), 862-877.
- Michels, Jeremy (2012), “Do Unverifiable Disclosures Matter? Evidence from Peer-to-Peer Lending,” *The Accounting Review*, 87 (4), 1385-1413.
- Mills, Gregory B. and William Monson (2013), “The Rising Use of Nonbank Credit among U.S. Households: 2009-2011,” *Urban Institute*.
- Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), “Mine Your Own

- Business: Market-Structure Surveillance through Text Mining,” *Marketing Science*, 31 (3), 521-543.
- Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards (2003), “Lying Words: Predicting Deception from Linguistic Styles,” *Personality and Social Psychology Bulletin*, 29 (5), 665-675.
- Nicholson, Nigel, Emma Soane, Mark Fenton-O’Creevy, and Paul Willman (2005), “Personality and Domain-Specific Risk Taking,” *Journal of Risk Research*, 8 (2), 157-176.
- Nyhus, Ellen K., and Paul Webley (2001), “The Role of Personality in Household Saving and Borrowing Behaviour,” *European Journal of Personality*, 15 (S1), S85-S103.
- Oppenheimer, Daniel M. (2006), “Consequences of Erudite Vernacular Utilized Irrespective of Necessity: Problems with Using Long Words Needlessly,” *Applied Cognitive Psychology*, 20 (2), 139-156.
- Ott, Myle, Claire Cardie, and Jeff Hancock (2012), “Estimating the Prevalence of Deception in Online Review Communities,” In *Proceedings of the 21st International Conference on World Wide Web*, 201-210.
- Pasupathi, Monisha (2007), “Telling and the Remembered Self: Linguistic Differences in Memories for Previously Disclosed and Previously Undisclosed Events,” *Memory*, 15 (3), 258-270.
- Pennebaker, James W. (1993), “Putting Stress into Words: Health, Linguistic, and Therapeutic Implications,” *Behaviour Research and Therapy*, 31 (6), 539-548.
- Pennebaker, James W., Roger J. Booth, and Martha E. Francis (2007), “Linguistic Inquiry and Word Count: LIWC [Computer Software],” *Austin, TX: liwc. net*.
- Pennebaker, James W., and Anna Graybeal (2001), “Patterns of Natural Language Use: Disclosure, Personality, and Social Integration,” *Current Directions in Psychological Science*, 10 (3), 90-93.
- Pennebaker, James W., and Laura A. King (1999), “Linguistic Styles: Language Use as an Individual Difference,” *Journal of Personality and Social Psychology*, 77 (6), 1296-1312.
- Pennebaker, James W., Tracy J. Mayne, and Martha E. Francis (1997), “Linguistic Predictors of Adaptive Bereavement,” *Journal of Personality and Social Psychology*, 72 (4), 863-871.
- Pennebaker, James W. and Lori D. Stone (2003), “Words of Wisdom: Language Use over the

- Life Span,” *Journal of Personality and Social Psychology*, 85 (2), 291-301.
- Pope, Devin G., and Justin R. Sydnor (2011), “What’s in a Picture? Evidence of Discrimination from Prosper.com,” *Journal of Human Resources*, 46 (1), 53-92.
- Preotiuc-Pietro, Daniel, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar (2015), “The Role of Personality, Age and Gender in Tweeting about Mental Illnesses,” *NAACL HLT*, 21-30.
- Ravina, Enrichetta (2012), “Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets,” *Available at SSRN*.
- Rugh, Jacob S., and Douglas S. Massey (2010), “Racial Segregation and the American Foreclosure Crisis,” *American Sociological Review*, 75 (5), 629-651.
- Shah, Anuj K., Sendhil Mullainathan, and Eldar Shafir (2012), “Some Consequences of Having too Little,” *Science*, 338 (6107), 682-685.
- Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michael Kosiniki, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar (2013), “Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach,” *Plos One*, 8 (9), e73791.
- Sievert, Carson, and Kenneth E. Shirley (2014), "LDavis: A method for Visualizing and Interpreting Topics." *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 63-70.
- Slatcher, Richard B., and James W. Pennebaker (2006), “How Do I Love Thee? Let Me Count The Words,” *Psychological Science*, 17 (8), 660-664.
- Sonenshein, Scott, Michal Herzstein, and Utpal M. Dholakia (2011), “How Accounts Shape Lending Decisions Through Fostering Perceived Trustworthiness,” *Organizational Behavior and Human Decision Processes*, 115 (1), 69-84.
- Tausczik, Yla R., and James W. Pennebaker (2010), “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods,” *Journal of Language and Social Psychology*, 29 (1), 24-54.
- Tibshirani, Robert (1997), “The Lasso Method for Variable Selection in the Cox Model,” *Statistics in Medicine*, 16(4), 385–395.

- Thomas, Lyn C. (2000), "A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers," *International journal of forecasting*, 16 (2), 149-172.
- Tirunillai, Seshadri, and Gerard J. Tellis (2012), "Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance," *Marketing Science*, 31 (2) 198-215.
- Toma, Catalina L., and Jeffrey T. Hancock (2012), "What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles," *Journal of Communication*, 62 (1), 78-97.
- Toubia, Olivier and Oded Netzer (2016), "Idea Generation, Creativity and Prototypicality," *Marketing Science*, forthcoming.
- Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe (2010), "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment," *ICWSM*, 178-185.
- Ventrella, Jeffrey J. (2011), *Virtual Body Language: The History and Future of Avatars: How Nonverbal Expression is Evolving on the Internet*. ETC Press.
- Weiss, Brent and Robert S. Feldman (2006), "Looking Good and Lying to Do It: Deception as an Impression Management Strategy in Job Interviews," *Journal of Applied Social Psychology*, 36(4), 1070-1086.
- Yarkoni, Tal (2010a), "Personality in 100,000 Words: A Large-Scale Analysis of Personality and Word Use among Bloggers," *Journal of Research in Personality*, 44 (3), 363-373.
- Yarkoni, Tal (2010b), "The Abbreviation of Personality, or How to Measure 200 Personality Scales with 200 Items," *Journal of Research in Personality*, 44 (2), 180-198.
- Zhang, Juanjuan, and Peng Liu (2012), "Rational Herding in Microloan Markets," *Management Science*, 58 (5), 892-912.

Table 1. Descriptive statistics for the Prosper data

Variables	Min	Max	Mean	SD	Freq.
Amount requested	1,000	25,000	6,507.3	5,732.9	
Debt-to-income ratio	0	10.01	.33	.89	
Lender interest rate	0	.350	.180	.077	
Number of words in description	1	766	207.9	137.4	
Number of words in title	0	13	4.593	2.015	
% of long words (6+ letters)	0	0.714	0.298	0.064	
SMOG	3.129	12	11.347	1.045	
Enchant Spellchecker	0	56	2.986	3.074	
# Prior Listings	0	67	2.016	3.097	
Credit grade: AA					0.086
A					0.082
B					0.186
C					0.219
D					0.170
E					0.128
HR					0.129
Loan repayment (1 = paid, 0 = defaulted)					0.669
Loan image dummy					0.670
Home owner dummy					0.470

Table 2. Area under the curve (AUC) for models with text only, financial and demographics information only, and a combination of both

	(1) Text only	(2) Financial/demog.	(3) Text & Financial/demog.	Improvement from (2) to (3)
Low credit grades: D, E, HR	61.33%	62.44%	64.96%	4.03%
Medium credit grades: B, C	62.37%	65.72%	68.13%	3.67%**
High credit grades: AA, A	71.31%	76.05%	78.09%	2.68%**
Overall AUC	66.68%	70.52%	72.56%	2.89%**
Jaccard Index	35.85%.	37.85%	38.85%	2.64%**

Notes: all AUCs reported in this table are the averaged across 10 replications of 10-folds mean. See Figure 1 for a plot of the receiver operating characteristic (ROC) curve for the average across a representative 10 fold. The Jaccard index is calculated as $N00/(N01+N10+N00)$, where N00 is the number of correctly predicted defaults, N01 and N10 are the numbers of mispredicted repayments and defaults, respectively. ** represents significant improvements at the 0.05 level.

Table 3. The thirteen LDA topics and representative words with highest relevance

Group	LDA topic	Words with highest relevance ($\lambda = 0.5$)
Purpose of the loan	Relocation Loan	Move, Live, New, Apartment
	School Loan	School, College, Student, Employ, Full, Time, Graduate
	Rate Reduction	Debt, Card, Interest, Rate, High, Credit, Consolidate
	Business Loan	Business, Company, Service, Base
Life circumstances	Prior Loan Details	Prosper, Thanks, List, Borrow, Lender
	Family Medical Issues	Bill, Husband, Wife, Medic, Family, Care
	Monthly Income	Month, Pay, Per, Every, Paid, Payday
	Housing	Home, Property, Purchase, Rental, Real, Invest
	Credit Score Details	Score, Credit, Account, Report, Year, Delinquency
	Monthly Expenses 1	Expense, Monthly, Household, Cloth
	Monthly Expenses 2	Payment, Car, Rent, Monthly
Pleading to lenders	Help, 2 nd Chance	Get, Help, Would Want, Try, Need
	Explanations	This, Good, Loan, Because, Candidate, Situation, Purpose

Note: the sample words are chosen based on the relevance measure with $\lambda = 0.5$. See Table A5 in the Web Appendix for list of words with top relevance in each topic.

Table 4. Binary regression with the thirteen LDA topics (repayment = 1)*

Financial and loan related variables	Estimate (Std. E)	Textual Variable	Estimate (Std. E)
Amount Requested (in \$10 ⁵)	-7.49 (0.37)	Number of words in Description (in 10 ⁴)	-5.77 (2.04)
Credit Grade HR	-0.85 (0.08)	Number of spelling mistakes	0.00 (0.01)
Credit Grade E	-0.50 (0.08)	SMOG (in 10 ³)	-2.99 20.9
Credit Grade D	-0.36 (0.06)	Words with 6 letters or more	-1.07 (0.45)
Credit Grade C	-0.21 (0.06)	Number of words in the title (in 10 ³)	-3.23 (8.65)
Credit Grade A	0.85 (0.08)	Prior Loan Details	0.18 (0.34)
Credit Grade AA	0.29 (0.07)	Relocation/Moving Loan	2.20 (0.46)
Debt To Income	-0.09 (0.02)	Rate Reduction	2.24 (0.35)
Images	0.05 (0.04)	College Loans	1.70 (0.37)
Home Owner Status	-0.29 (0.04)	Business Loan	-0.51 (0.26)
Lender Interest Rate	-5.04 (0.32)	Credit Score details	1.66 (0.39)
Bank Draft Fee Annual Rate	-33.13 (19.44)	Monthly Income	0.71 (0.43)
Prior Listings	-0.03 (0.01)	Housing details	0.30 (0.36)
Intercept	2.89 (0.39)	Family Medical Issues	-2.98 (0.40)
		Hardworking-Responsible	-0.73 (0.37)
		Help 2nd Chance	-0.78 (0.40)
		Monthly expenses 2	0.89 (0.40)

* Bold face for P-value ≤ 0.05 . For brevity we do not report in this table the estimates of the demographics variables such as location, age, gender and race.

Table 5. Binary regression with LIWC (repayment = 1)*

Variable	Beta (Std. E)	Variable	Beta (Std. E)	Variable	Beta (Std. E)	Variable	Beta (Std. E)
Financial and basic text variables:		LIWC dictionary:					
Amount Requested(x 10⁵)	-7.163 (0.3668)	Swear words	35.5112 (35.275)	Past words	-2.1032 (1.9895)	Person pronoun words	0.4119 (6.6093)
Credit Grade HR	-0.8551 (0.0844)	Filler words	13.3939 (6.224)	Inhibition words	-2.3047 (3.4172)	Work words	0.5175 (0.9333)
Credit Grade E	-0.4642 (0.0817)	Perception words	13.4328 (10.839)	Home words	-2.3822 (1.7643)	Sexual words	-10.5097 (10.828)
Credit Grade D	-0.3383 (0.0623)	Relative words	9.1729 (2.3748)	Hear words	-2.4191 (14.038)	They words	-15.491 (9.3357)
Credit Grade C	-0.1959 (0.0559)	Friend words	9.7894 (7.0217)	I words	-2.7392 (8.1836)	Positive emotion words	0.2869 (2.0477)
Credit Grade A	0.7837 (0.0802)	Anxiety words	8.7494 (8.9305)	Tentative words	-2.8712 (2.0522)	Money words	0.2085 (0.7944)
Credit Grade AA	0.2838 (0.0692)	Negate words	6.0709 (3.3228)	Non-fluency words	-3.2295 (9.518)	Ingest words	-0.0434 (5.279)
Debt To Income	-0.0906 (0.0186)	Insight words	5.0732 (2.8214)	Anger words	-3.2911 (9.7405)	Verbs words	-0.1936 (1.3174)
Images	0.0599 (0.0389)	We words	4.1277 (8.3628)	Achieve words	-3.3204 (1.5601)	Adverbs words	-0.3578 (1.8814)
Home Owner Status	-0.3199 (0.0381)	Pronoun words	3.7935 (9.9981)	Incline words	-3.5433 (2.3316)	Functional words	-0.9427 (1.8725)
Lender Interest Rate	-5.2556 (0.3148)	Exclusion words	3.1073 (2.7497)	She/he words	-3.5689 (7.3598)	Bios words	-1.3376 (2.6575)
Bank Draft Fee Annual Rate	-33.9126 (19.509)	Sad words	2.9955 (6.4981)	You words	-3.714 (8.8219)	Assent words	-1.3651 (14.463)
Prior Listings	-0.0236 (0.0058)	Quantitative words	2.7495 (1.9363)	Cause words	-3.7248 (2.407)	Family words	-1.4804 (2.8298)
Number of words in Description(x 10 ⁴)	-3.494 (1.96)	Articles	2.457 (2.0896)	Social words	-4.2882 (1.5697)	I pronoun words	-1.72 (10.026)
Number of spelling mistakes	-0.0124 (0.0068)	Numbers words	2.2907 (2.7328)	Health words	-4.7602 (4.2679)	Death words	-16.3445 (10.721)
SMOG	-0.0252 (0.0209)	Preposition words	2.1719 (1.8415)	Certain words	-5.2433 (2.7262)	Body words	-19.1156 (5.8326)
Words with 6 letters or more	0.4455 (0.5716)	Conjoint words	1.8673 (1.8392)	Present words	-6.223 (1.7067)	Religion words	-20.2865 (6.7741)
Number of words in the title	-0.0062 (0.6035)	Auxiliary verbs words	1.7732 (2.3818)	Human words	-7.5781 (3.5803)	Feel words	-24.617 (11.734)
(Intercept)	3.6557 (0.6035)	Affect words	1.2929 (1.5234)	Space words	-8.2317 (2.5648)	See words	-10.5021 (11.597)
		Discrepancy words	1.2769 (2.686)	Future words	-8.4576 (3.5391)	Leisure words	0.5548 (2.6577)
		Cognitive mechanism words	0.7625 (1.8828)	Motion words	-9.1849 (2.8071)		
		Negative emotion words	0.7453 (4.7407)	Time words	-9.4218 (2.3077)		

* Bold face for P-value ≤ 0.05. For brevity we do not report in this table the estimates of the demographics variables such as location, age, gender and race.

Figure 1. Receiver operating characteristics (ROC) curves for models with text only, financial and demographics information only, and a combination of both

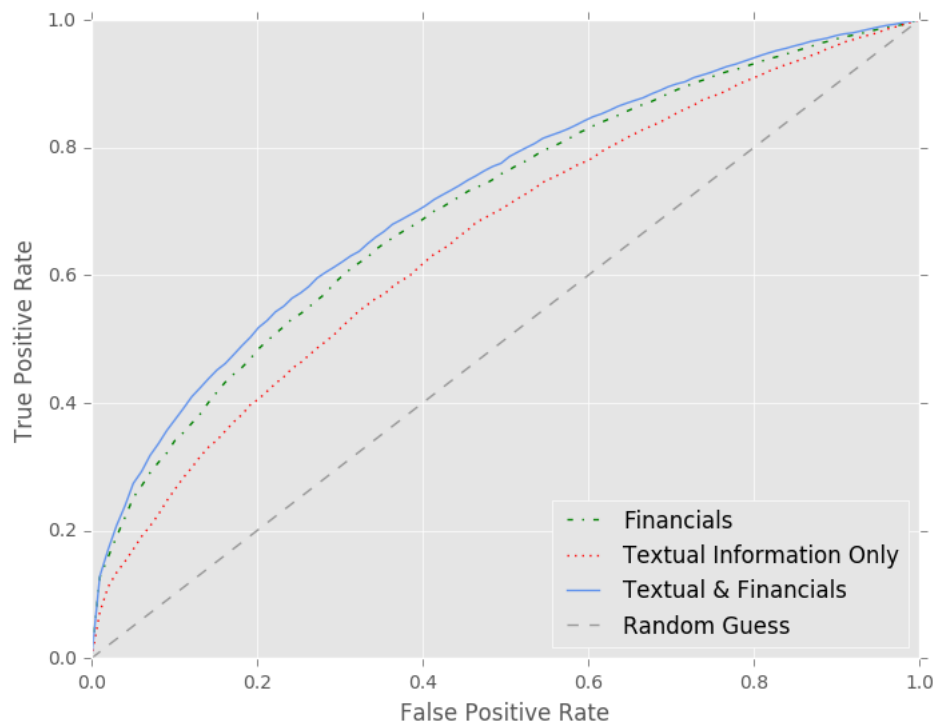
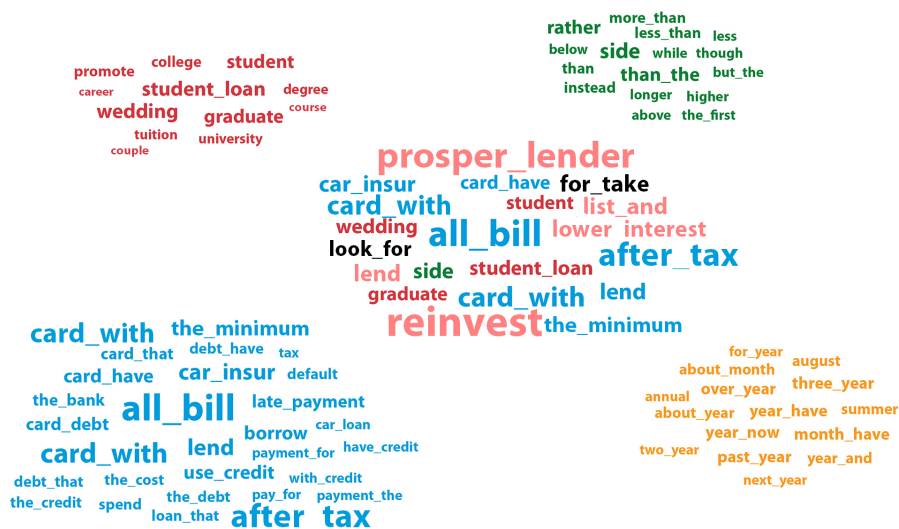


Figure 2. Words indicative of paying back the loan



Note: The most common words appear in the middle cloud (cutoff = 1:1.5) and then organized by themes. On the top, in green, and clockwise: relative words, time related words, words related to borrowing and debt, and words related to a brighter financial future.

[illegible]

Figure 4: LDA analysis – selecting the number of topics, out-of-sample AUC

