When Words Sweat:

Identifying Signals for Loan Default in the Text of Loan Applications

Oded Netzer
Professor of Business
Columbia Business School
Columbia University
onetzer@gsb.columbia.edu

Alain Lemaire
Doctoral student
Columbia Business School
Columbia University
alemaire18@gsb.columbia.edu

Michal Herzenstein
Associate Professor of Marketing
Lerner College of Business and Economics
University of Delaware
michalh@udel.edu

May, 2019

Equal authorship. Financial support from Columbia Business School and Lerner College at the University of Delaware is greatly appreciated.

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 120 thousand loan requests from Prosper.com, an online crowdfunding platform. Including the textual information in the loan significantly helps predict loan default and can have substantial financial implications. The authors find that loan requests written by defaulting borrowers are more likely to include words related to their family, mentions of god, the borrower's financial and general hardship, pleading lenders for help, and short-term focused words. The authors further observe that defaulting loan requests are written in a manner consistent with the writing style of extroverts and liars.

Keywords: consumer finance, loan default, text mining, machine learning

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, 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." Which borrower is more likely to default? This question is at the center of our research, as we investigate the power of words in predicting loan default. We claim and show that the text borrowers write at loan origination provides valuable information that cannot be otherwise extracted from the typical data lenders have on borrowers (which mostly include financial and demographic data), and that additional information is crucial to predictions of default. The idea that text written by borrowers can predict their loan default, builds on recent research showing that text is indicative of people's psychological states, traits, opinions, and situations (e.g., Humphreys and Jen-Hui Wang 2018; Matz and Netzer 2017).

In essence, the decision whether or not to grant a loan depends on the lender's assessment of the borrower's ability to repay it. But this assessment is often difficult because loans are repaid over a lengthy period of time, during which unforeseen circumstances may arise. For that reason, lenders (e.g., banks) and researchers collect and process as many pieces of information as possible within this tightly regulated industry[1]. These pieces of information can be classified into four categories: (1) The borrower's financial strength, which is reflected by one's credit history,

[1] Indeed, our personal discussions with data scientists at banks and other lending institutions suggest that they analyze as many as several thousand data points for each borrower.

FICO score, income, and debt (Mayer, Pence, and Sherlund 2009), is most telling of the borrower's ability to repay (Avery et al. 2000); (2) Demographics, such as gender or geographic location (Rugh and Massey 2010);[2] (3) Information related to the loan—the requested borrowed amount and interest rate (Gross et al. 2009); (4) Everything else that can be learned from interactions between borrowers and people at loan granting institutions. Indeed, Agarwal and Hauswald (2010) found that supplementing the loan application process with the human touch of loan officers significantly decreases default rate due to better screening and higher interpersonal commitment from borrowers. However, these interactions are often laborious and expensive.

Indeed recently, human interactions between borrowers and lenders have been largely replaced by online lending platforms operated by banks, other lending institutions, or crowdfunding platforms. In such environments, the role of both hard and soft pieces of information becomes crucial. Accordingly, our main proposition is that the text people write when requesting an online crowdfunded loan provides additional important information on the borrower, such as their intentions, personality, and circumstances—information that cannot be deduced from the financial and demographic data and in a sense analogous to body (or unspoken) language detected by loan officers. Furthermore, because the evaluation of borrowers by loan officers has been shown to provide additional information over and beyond the financial information (Agarwal and Hauswald 2010) we contend that in a similar vein, the text borrowers write when applying for a crowdfunded loan is predictive of loan default above and beyond all other available information. This hypothesis extends the idea that our demeanor can be a manifestation of our true intentions (DePaulo et al. 2003) into the text we write.

To answer the question we posed earlier—who is more likely to default—we apply text-

[2] Following the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act (FHA) most demographic variables cannot be used directly in the decision to grant loans.

mining and machine learning tools to a dataset of over 120 thousand loan requests from the crowdfunding platform Prosper. Using an ensemble stacking approach that includes tree-based methods and regularized logistic regressions, we find that the textual information significantly improves predictions of default. A simple back-of-an-envelope analysis shows that using textual information can increase lenders' ROI over an approach that uses only financial and demographics information by as much as 5.75%.

To learn which words, writing styles, and general ideas conveyed by the text are more likely to be associated with defaulted loan requests we further analyzed the data using a multi-method approach including both machine learning tools (e.g., naïve Bayes and L1 regularization binary logistic model), as well as standard econometric tools such as logistic regression of the topics extracted from a latent Dirichlet analysis (LDA) analysis, and the sub-dictionaries of the Linguistic Inquiry and Word Count dictionary (LIWC; Tausczik and Pennebaker 2010). Results across the first three analyses consistently show that loan requests written by defaulting borrowers are more likely to include words (or themes) related to the borrower's family, financial and general hardship, mentions of god, and the near future, as well as pleading lenders for help, and using verbs in present and future tenses. Therefore, the text and writing style of borrower #1 in our opening example suggest this person is more likely to default. In fact, all else equal, our analysis shows that based on the loan request text, borrower #1 is approximately eight times more likely to default relative to borrower #2. These analyses demonstrate the successful use of machine learning tools in going beyond merely predicting the outcome, and into the realm of interpretation, by inferring the words, topics, and writing styles that are most associated with a behavioral outcome.

Applying the LIWC dictionary to our data allows a deeper exploration into the potential

traits and states of borrowers. Our results suggest that defaulting loan requests are written in a manner consistent with the writing styles of extroverts and liars. We do not claim that defaulting borrowers were intentionally deceptive when they wrote the loan request; rather, we believe their writing style may have reflected, intentionally or not, doubts in their ability to repay the loan).

Our examination into the manifestation of consumers' personalities and states in the text they write during loan origination contributes to the fast-growing literature in consumer financial decision making. Consumer researchers have been investigating factors that affect consumer saving (Dholakia et al. 2016, Sussman and O'brien 2016), debt acquisition, and repayment (Herzenstein, Sonenshein, and Dholakia 2011). Most of these investigations have been done on a smaller scale, such as with experimental participants or smaller datasets. We add to this literature by showing, on a large scale and with archival data, how participants in the crowdfunding industry can better assess the risk of default by interpreting and incorporating soft unverifiable data (the words borrowers write) into their analysis.

The rest of this paper is organized as follows. In the next section we discuss the limitations of structured financial and demographic data in the context of consumer financial decisions and the opportunity in leveraging textual information. Specifically, drawing on extant literature we argue that credit scores and demographics miss important information about borrowers that lenders can learn from mining the text these borrowers write. We then delineate the data, our text-mining and modeling approaches, results, and their generalization.

WHAT DO HARD DATA MISS? THE OPPORTUNITY IN TEXT

Financial data such as credit scores have been used extensively to predict consumers' credit riskiness, which is the likelihood they will become seriously delinquent on their credit obligations over the next two years. FICO, the dominant credit scoring model, is used by 90% of

all financial institutions in the U.S. in their decision-making process (according to myFICO.com). In calculating consumers' scores, credit score agencies take into account a multitude of data points including past and current loans, credit cards—their allowance and utilization, as well as any delinquencies that were reported to the credit bureau (by companies such as cable TV, cell phone providers etc.). However, while the usage of credit scores and financial information clearly has its benefits in predicting consumer risk, these measures have been found to be insufficient and even biased. For example, Avery et al. (2000) argue that credit scores are snapshots of the past because they rely on available consumers' credit history, and therefore miss important factors such as health status and length of employment which are more forward looking in nature. The authors find that those with low credit scores often benefit most from such additional data. Agarwal, Skiba, and Tobacman (2009) provide further evidence for this deficiency of credit scores by showing that the more flexible credit score systems, which weigh differently the different data sources that comprise the score for each person, are eight times better at predicting loan default than the more rigid systems such as credit scores. In sum, credit scores might predict future financial behavior well for some but not others.

Analyzing the subprime mortgage default during 2006-7, Palmer (2015) and Sengupta and Bhardwaj (2015) show that current contextual information such as loan characteristics (e.g., loan amount) and economic characteristics (e.g., housing prices) are often more predictive of mortgage default than traditional financial measures such as credit scores and debt-to-income ratios. Thus, it is clear that credit scores are missing important information that may be predictive of consumers' financial health and that the heavy reliance on such scores might blind lending organizations from looking at additional sources of information.

Recognizing the limitation of financial measures, financial institutions and researchers

added other variables to their predictive models, mostly related to the individuals' demographic, geographic, and psychographic characteristics (e.g., Barber and Odean 2001; Pompian and Longo 2004). However, other characteristics, such as those related to emotional and mental states, as well as personalities, have been found to be closely tied to financial behaviors and outcomes (Norvilitis et al. 2006; Rustichini et al. 2016), yet are still missing from those predictive models. The problem of reliance on purely structured financial information is even more severe in the growing realm of crowdfunded unsecured loans, because human interactions between lenders and borrowers is scarce. But, therein lies the problem. Personalities and mental states are difficult, if not impossible, to infer from financial data alone, thus the pressing need to go beyond such traditional data when attempting to predict behavior. We suggest and demonstrate that the text borrowers write at loan origination can provide the much needed supplemental information. Specifically, this textual information can be useful in understanding not only the past behavior of the borrower, but also the present context of the loan and the borrower's future intentions.

LANGUAGE STYLES, INDIVIDUAL CHARACTERISTICS, AND FINANCIAL BEHAVIORS

Our proposition that text can predict default builds on research in marketing, psychology, linguistics, and finance that establishes two connections: (1) word usage and writing styles are indicative of some stable inner traits as well as more transient states; and (2) these traits and states affect people's financial behaviors. In this section we provide evidence for these two connections.

We begin with the relationship between words and personality traits, emotional states, and demographics. The premise that the text may be indicative of deeper traits is predicated on the idea that there is a systematic relationship between the words people use and their personality traits (Hirsh and Peterson 2009) 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), blogs (Yarkoni 2010), social media (Schwartz et al. 2013), and daily speeches (Mehl, Gosling, and Pennebaker 2006). 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. Therefore, even if the content might be similar across different individuals, the manner in which they convey that content differs.

Research over the last two decades established the association between word usage with the text originator's personality, focusing on the big five personality traits—extroversion, agreeableness, conscientiousness, neuroticism, and openness (Kosinski, Stillwell, and Graepel 2013; Pennebaker and Graybeal 2001; Pennebaker and King 1999). Using a sample of over 1,000 students, Pennebaker and King (1999) collected scales for the big five personality traits, emotional states (PANAS), as well as two essays related to their current thoughts and feelings and to coming to college. Using LIWC and correlational analyses, the authors concluded that extroversion is associated with more positive emotions words usage, neuroticism with more first person singular and with negative emotions words, openness is associated with more articles usage while agreeableness with fewer articles, and conscientiousness is associated with fewer negations. These results were later corroborated and extended by many, see specifically Yarkoni (2010) and Schwartz et al. (2013; Figure 2). While some of the aforementioned results are expected (especially those around emotions), some are not—why would people who are opened to experiences use more articles? According to Yarkoni, this results reflects "a fundamental difference in language style rather than content… suggesting a potential tendency to favor high-frequency function words at the expense of the lower frequency content words" (page 368). Language style, more so than content, is at the heart of our research.

In addition to personality traits words have been found to be associated with one's emotional and behavioral states (Pennebaker, Mayne, and Francis 1997) such as physical and mental health (Preotiuc-Pietro et al. 2015), impulsivity (Arya, Eckel, and Wichman 2013), and deception (Newman et al. 2003; Ott, Cardie, and Hancock 2012) among others. Text has also been use to predict demographic characteristics such as age, gender, education, and income (Pennebaker and Stone 2003; Schwartz et al. 2013). For example, frequent usage of long words (6 letters or more) was found to be related to higher educational levels (Tausczik and Pennebaker 2010), and spelling mistakes in written text are correlated with income levels (Harkness 2016).

Next we examine the relationship between personality traits, emotional states and financial behavior. The aforementioned literature conveys the potential in analyzing text to extract a wealth of information otherwise unobtainable in many settings. Such information has been shown to be a valuable predictor of financial behaviors. For example, Anderson et al. (2011) show that credit scores are correlated with the big five personality traits. They find negative correlations between extroversion and conscientiousness and credit scores. Extroverts wish to have exciting life styles and hence sometimes spend beyond their means. The result for conscientiousness is a bit surprising because these are diligent and responsible people, however their need for achievement is high which might induce a pattern of spending that is larger than their means. Berneth, Taylor, and Walker (2011) found that agreeableness is negatively correlated with credit scores because these people aim to please and are less likely to say no to unnecessary expenses. The extent to which such traits are predictive of financial behavior over and beyond credit scores is an empirical question, which we investigate in this research.

Other individual characteristics are also related to financial behavior. Arya, Eckel, and Wichman (2013) find that credit scores are correlated with personality, time and risk preference,

trustworthiness, and impulsiveness; Nyhus and Webley (2001) show that emotional stability, autonomy, and extroversion are robust predictors of saving and borrowing behaviors; and Norvilitis et al. (2006) show that debt is related to delay of gratification. Financial behaviors such as saving, taking loans, and credit card usage, as well as overall scores such as FICO have also been found to be correlated with education (Nyhus and Webley 2001), financial literacy (Fernandes, Lynch, and Netemeyer 2014), number of hours people work (Berneth, Taylor, and Walker 2011), stress and physical wellbeing (Netemeyer et al. 2018), self-regulation (Freitas et al. 2002), and even the number of social media connections (Wei et al. 2015).

In sum, we postulate that many of the behavioral characteristics and situational factors that have been found to be related to financial behaviors, such as personality traits and future intentions, can be extracted from the text borrowers write in their loan request. Therefore, that text is an invaluable addition to the hard financial data when predicting loan default.

SETTINGS AND DATA

We examine the value of text in predicting 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 $14 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. We downloaded all loan requests posted between April 2007 and October 2008, a total of 122,479 listings. 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. We chose data from the earlier days because it is richer and more diverse particularly with respect to the textual information in the loan requests.

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). Table A1 in the Web Appendix[3] presents correspondence between Prosper's credit grades and FICO score.[4] 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 write in that textbox are at the center of our research.

Borrowers could choose to become part of a group of borrowers. These groups often have a subject matter (groups can be about alma mater, certain purpose for the loan, geographical regions, certain professions, etc.) and they must have a leader that supposedly acts as another layer of vetting (group leaders at the time received $12 from Prosper for each group member whose loan request is funded, so they have an incentive to enlarge their group). Group membership may affect the likelihood of a loan being granted and its default likelihood (Hildebrand, Puri, and Rocholl 2017). Accordingly, group affiliation is include in our models. After borrowers posted their listings, they may interact with lenders via a Q&A interface. Unfortunately, we do not have access to those chats, but unofficial reports suggest that the Q&A feature usage was very limited (3% of the lenders participated).

Since our interest is in predicting default, we focus on those loan requests that were

[3] All tables and graphs whose numbering begins with "A" are presented in the web appendix.
[4] The distribution of credit scores among Prosper borrowers in our sample is different from the distribution in the U.S. at the time (see Table A2): We have fewer people with high credit scores and more with medium scores.

funded (19,446 requests). The default rate in our sample is 35%.

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 excluded from our analysis numbers and symbols).[5] 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)[6]. To reduce the dimensionality of the textual data and avoid over-relying on more obscure words, we focus our analyses on the most frequent stemmed words and bi-grams that appeared in at least 400 loan requests, which left us with 1,052 bi-grams.[7]

Textual, Financial, and Demographic Variables

Our dependent variable is loan repayment/default as reported by Prosper[8] (binary: 1 = paid in full, 0 = defaulted). Our data horizon ends in 2008, and all Prosper loans at the time were to be repaid over three years or less, therefore we know whether each loan in our database was repaid or defaulted—there are no other options. The set of independent variables we use includes textual, financial, and demographic variables. We elaborate on each group next.

Textual variables. These 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,

[5] Because of stemming, words with less than 3 letters such as "I" may be kept due to longer stems (e.g., I've).
[6] While n-grams with n > 2 (such as strings of three or more words) could have been part of our analyses, it increases dimensionality and computational difficulty substantially, which ultimately precluded their inclusion.
[7] We note that our analyses are robust to increasing the number of words and bi-grams that are included.
[8] 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."

however the evidence is inconclusive. Hancock et al. (2007) show 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 indicate 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, suggesting the higher language was likely used deceptively (Oppenheimer 2006). (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 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 because it serves as a proxy for characteristics correlated with lower income. (5) The 1,052 *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 m in loan request j is defined by $tf_{mj} = X_{mj}/N_j$, where X_{mj} is the number of times

word m appears in loan request j, and N_j is the number of words in loan request j. This component controls for the length of the document. The inverse-document-frequency is defined by $idf_m = log(D/M_m)$, where D is the number of loan requests and M_m is the number of loan requests in which word m appears. This terms controls for the how often a word appears across documents. *Tf-idf* is given by: $tf - idf_{mj} = tf_{mj} \times (idf_m + 1)$. Taken together, the tf-idf statistic provides a measure of how likely a word is to appear in a document over and beyond chance.

Financial and Demographic Variables. The second type of variables we consider are financial and demographic information, commonly used in traditional risk models. These include all information available to lenders on Prosper—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. We also control for the geographical location of the borrower to account for differences in the economic environment of the borrower. We grouped the borrowers' states of residency into eight groups based on the Bureau of Economic Analysis classification (plus a group for Military personnel serving overseas). Finally, in order to fully account for all the information in loan requests, we extracted information included in the borrower's profile pictures, such as gender (Male, Female, and "cannot tell"), age brackets (Young, Middle-aged, Old), and race (Caucasian, African American, Asian, Hispanic, or "cannot tell") using human coders. See Web Appendix for details about the coding procedure.[9] We note that while the Equal Credit Opportunity Act forbids discrimination based on race, age, and gender, we include these variables in the statistical model in order to

[9] Our judges also rated borrowers' attractiveness and trustworthiness based on their picture (Pope and Sydnor 2011), but given the high degree of disagreements across raters, they are not included in our analyses.

examine the marginal value of the textual information that does not relate to demographics.

Lastly, we include the final interest rate for each loan as a predictor in our model.[10] 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, which means that including the final interest rate in the models should render other variables insignificant predictors. However, given the possibility that Prosper's bidding mechanism allows for some strategic behavior by sophisticated lenders (thus not fully reflecting a market efficient behavior), our models test whether the text is predictive over and beyond the final interest rate. Table 1 presents summary statistics for the variables in our model.

*** Insert Table 1 about here ***

PREDICTING DEFAULT

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 origination. In order to do so, we need to first build a strong benchmark—a powerful predictive model that includes the 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). Given the large number of predictors, and the predictive nature of the task at hand, machine learning methods are most appropriate. In the subsequent section, as we aim to understand *which words* predict default, we combine the machine learning methods with data reductions methods (e.g., topic modeling) and standard econometric tools. 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

[10] An alternative measure would be the maximum interest rate proposed by the borrower. However, because this measure is highly correlated the final lender rate, we include only the latter in the model.

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 choose the best model, create an ensemble or stack the different models. Recent research has demonstrated the superior performance of ensemble models relative to individual models (Lee, Hosanagar, and Nair 2018). An ensemble of models benefits from the strength of each individual model and, at the same time, reduces the variance of the prediction.

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 estimate five types of models in the first step. The models vary in terms of the classifier used and the approach to model variable selection. The five models are described below and include two logistic regressions and three versions of decision tree classifiers.[11]

Regularized Logistic Regressions. We estimate two logistic regressions—L1 and L2 regularization logistic regressions. The penalized logistic regression likelihood is:

$$L(Y|\beta,\lambda) = \sum_{j=1}^{n}(y_j \log\left(p\left(X_j|\beta\right)\right) + (1-y_j)\log\left(1-p\left(X_j|\beta\right)\right) - \lambda J(\beta),$$

where $Y = \{y_1, \dots, y_n\}$ is the set of binary outcome variables for n loans (loan repayment), $p(X_j|\beta)$ is the probability of repayment based on the logit model, where X_t is a vector of textual, financial, and demographic predictors for loan j, β are a set of predictors' coefficients, λ is a tuning penalization parameter to be estimated using cross-validation on the calibration sample, and $J(\beta)$ is the penalization term. The L1 and L2 models differ with respect to the functional

[11] We also considered another decision tree (AdaBoost) as well as a Support Vector Machine classifier but dropped them due to poor performance on our data.

form of the penalization term, $J(\beta)$. In L1, $J(\beta) = \sum_{i=1}^{k} |\beta_i|$, while in L2, $J(\beta) = \sum_{i=1}^{k} {\beta_i}^2$, where k is the number of predictors. Therefore, L1 is the Lasso regression penalty (shrinks many of the regression parameters to exactly zero), and L2 is the ridge regression penalty (shrinks many parameters to small but non-zero values). Before entering the variables into these regression we standardize all variables (Tibshirani 1997).

Tree-based Methods (Random forest and Extra Trees). There are three tree-based methods in the ensemble. We estimate two different Random Forest models, one with variance selection and the second with best feature selection as well as Extremely Randomized Trees (Extra Trees). Both models combine many decision trees, thus, each of these tree-based methods is an ensemble in and of itself. 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. In the Variance Selection Random Forest features are chosen based on a variance threshold determined by cross validation (80/20 split). In the K-Best Feature Selection Random Forest features are selected based on a χ^2 test. That is, we select the K-features with the highest χ^2 score. We use cross-validation (80/20 split) to determine the value of K. The Random Forest approach mitigates the problem of over-fitting in traditional decision trees. The Extra Trees is an extension of the Random Forest in which the cut-off point (the split) for each feature in the tree are also chosen at random (from a uniform distribution) and the best split among them is chosen. Due to the size of the feature space, we first apply a K-Best Feature Selection as described above to select the features to be included in the Extra Trees (see Web Appendix for more details).

We used the scikit learn package in Python (http://scikit-learn.org/) to implement the five 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). We use a randomized parameter optimization rather than an exhaustive search (or a grid search) due to the large number of variables in our model. The parameters are sampled from a distribution (uniform) over all possible parameter values.

Model Stacking and Predictions. In the second step, we estimate the weights for each model to combine the ensemble of models using the remaining 20% of the calibration data. We use a simple binary logistic model to combine the different predictive models. Though other classifiers may be used, a logistic binary regression meta-classifier helps avoid overfitting and often results in superior performance (Whalen and Gaurav 2013). In our binary logistic regression model, repayment is the dependent variable and the probabilities of repayment for each loan by each of the five models in the ensembles from step one (the two logistic regularization regressions and the three decision trees methods) as predictors. The estimated parameters of the logistic regression provide the weights of each individual model in the ensemble. Specifically, the ensemble repayment probability for loan *j* can be written as:

$$p(repayment_j) = \exp(\boldsymbol{x_j}'\boldsymbol{w}) / (1 + \exp(\boldsymbol{x_j}'\boldsymbol{w})),$$

where $\boldsymbol{x_j}$ is the vector of repayment probabilities for each model *s*— $p(repayment_j | model_s)$ from step one, and $\boldsymbol{w}$ are the estimated weights of each model in the logistic regression classifier.

We estimated an ensemble of the aforementioned five models, and find the following weights for the different model: L1 = 0.040, L2 = 0.560, Random Forest K-Best = 0.218, Random Forest Variance-Select = 0.116, and Extra Trees = 0.066.

To test whether 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 we get a robust measure of 100 predictions. 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. We further report the Jaccard index (e.g., Netzer et al. 2012) of loan default, which is defined as the number of correctly predicted defaulting loans divided by the total number of loans that were defaulted but we incorrectly predicted to be repaid, loans that were predicted to default 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, building on research that shows credit scores have a lower predictive power of financial behavior for people with low scores (Avery et al. 2000), we report predictions for high (AA, A), medium (B, C), and low (D, E, HR) credit grades (while controlling for credit grades within each group).

We compare three versions of the ensemble: (1) a model calibrated only on the financial and demographic data; (2) a model that includes just the textual information (i.e., all the variables we created from the freely written text borrowers constructed) and ignores the financial and demographic information, and (3) a model that includes financial and demographic information together with the textual data. Comparing models (2) and (3) provides the incremental predictive power of the textual information over 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. Additionally, we estimated separately the five predictive models that contribute to the ensemble (L1 and L2 regularization logistic regressions, the two Random Forest models and the Extra Trees model), to assess the value of the textual information in different models.

Prediction Results

Table 2 details the average results of the 10-fold cross validation across 10 random shuffling of the observations. The results we present are clean *out-of-sample* validation because in each fold we calibrate feature selection, model estimates, and the ensemble weights on 90% of the data and leave the remaining 10% of the data for validation. The results in Table 2 are the Area Under the ROC Curve (or AUC) and the Jaccard Index prediction measures. Figure 1 depicts the average ROC curve with and without the textual data for one randomly chosen 10-fold cross validation. As the ROC curve gets closer to the upper left corner of the graph, the underlying model is better at predicting the outcome. The AUC of the model with textual, financial, and demographics information is 2.64% better than the AUC of the model with only financial and demographics information, and this difference is statically significant. In fact, the model with both textual and financial information has higher AUC in all 100 replications of the cross-validation exercise. Breaking the sample by credit grade, we note that the textual information significantly improves predictions across all credit grade levels. However, the textual information is particularly useful in improving default predictions for borrowers with low credit levels. This result is consistent with Avery et al. (2000) who find credit scores to be least telling of people's true financial situation for those with low scores.

Interestingly, if we were to ignore the financial and demographic information and use only the borrower textual information, we obtain an AUC of 66.69% compared to an AUC of

70.72% for the model with only financial and demographic information. That is, a brief, unverifiable, "cheap talk" (Farrell and Rabin 1996), textual information provided by borrowers is nearly as predictive as the traditional financial and demographic information. This result is particularly impressive given the tremendous effort and expenditure involved in collecting the financial information relative to the simple method used to collect the textual information. This result may also suggest that textual information may be particularly useful in "thin file" situations, where the financial information about consumers is sparse[12].

The bottom part of Table 2, presents the predictive performance of each of the individual model in ensemble. We see that for each of the models the textual information significantly improves predictions in the validation sample over and beyond the model with the financial and demographics information only. However, the stacking ensemble model further improve predictions over each of the independent models. There are two key takeaways from this comparison. First, that the textual information itself, independent of the machine learning model used, significantly contributes in predicting default over the traditional financial measures (Banko and Brill 2001). Second, that combining models using an ensemble learning model further helps in predicting default. The reason for the improvement of the ensemble learning model relative to the individual model is that different models perform better in different aspects of the data. To better understand the performance of the ensemble model relative to the individual models, we compared the performance of the individual models by credit score, or the frequency of the words across loans (see Table A3). We find, for example, that the Random Forest Variance and Extra Trees perform particularly well for low credit score loans, while the regularized logit models perform well for the high credit score loans. Similarly, we find that the

[12] https://www.cutimes.com/2017/09/22/the-thin-filed-underbanked-an-untapped-source-of-buyers/?slreturn=20190103095108

Random Forest Select model perform best for words that were less frequent across loans, but the regularized logit model perform best for more frequent words.

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

To quantify the managerial relevance and financial implications of the improvement in predictive ability offered by the textual data, we conducted a back-of-the-envelope calculation. For each of the 19,446 granted loans we calculated the expected profit from investing $1,000 in each loan based on the models with and without text. In calculating the expected profit, we assume that borrowers who default repay on average 25% of the loan before defaulting (based on estimates published by crowdfunding consulting agencies). The expected profits for loan *j* is:

$$E(profit_j) = \left[1 - Prob(repayment_j)\right] \times \left[-0.75 \times amount\_granted_j + 0.25 \times Interest\_earned\right] + \left[Prob(repayment_j)\right] \times Interest\_earned, \quad (1)$$

where $Prob(Repayment_j)$ is the probability of repayment of loan *j* based on the corresponding model (with or without the text), $amount\_granted_j$ is the principal amount the lender grants for loan *j* ($1,000 in our case), and $Interest\_earned_j$ is the interest rate paid to the lender based on loan j^{th} final interest rate, over three years for repaid loans and over three quarters of a year for defaulted loans (for simplicity we do not time-discount payments in years 2 and 3). For each of the two policies (based on the model with and without text) we sort the loans based on their expected profit and select the top 1,000 loans with the highest expected return for each policy. Finally, we calculate the *actual* profitability of each lending policy based on the actual default of each loan in the data to calculate the return on the investment on the million dollars (1,000 loans time $1,000 per loan). We find that the investment policy based on the model with the textual data returns $57,571 more than the policy based on the financial and demographics information only. This is an increase of 5.75% in the return on investment (ROI) on the million dollars invested. Thus, while the improvement in default prediction for the model with the textual information

might seem modest (nearly 3%) even though it is statistically significant, the improvement in ROI based on the textual information, is substantial and economically meaningful.

We note that because we only know the repayment outcome for funded loans and because our model was trained to predict default only on the sample of funded loans our back-of-the-envelope can only assesses the benefit of funding loans for loans that were funded. However, our model and data can provide some predictions with respect to the default likelihood of rejected (unfunded) loans. Indeed, when we compare the default likelihood distribution of rejected versus funded loans, we see that the default distribution, based on our model, is much higher for rejected loans relative to funded loans, but there is also a substantial overlap (see Figure A1). We also calculated a confusion matrix (see Table A4) based on the predicated default and expected profitability (following Equation 1), for loans that were actually granted versus not-granted and whether these loans should have been granted (generate positive return). Of the 19,446 funded loan requests, our model recommends funding 60% (11,795 loans). In addition, of the 103,033 unfunded loan requests, our model recommends funding 21% (21,631 loans). Overall, based on our model we recommend granting 33,426 loans that are predicted to generate positive profits.

To summarize, the text borrowers write in their loan request can significantly improve predictions of loan default over and beyond all other available information, including the loan's interest rate. The ensemble-based predictive model was chosen to maximize predictive ability, but it provides little to no interpretation of the parameter estimates, words, and topics that predict default. In the second part of the paper, we demonstrate how machine learning approaches combined with econometric models can be used beyond predictions and towards understanding of which words and writing styles are 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 all other information is perhaps surprising, given borrowers can write whatever they want. 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 (Hirsh and Peterson 2009). We employed four approaches to uncover whether words, topics, and writing styles of defaulters differ from those who repaid their loan (based on the sample of 19,446 funded loans). (1) We use a naïve Bayes classifier to identify the words or bi-grams 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 hence cannot control for other variables. (2) To alleviate this concern, we use a logistic regression with L1 penalization, which reduces the dimensionality of the word space by setting some of the parameters to zero, to uncover the words and bi-grams that are associated with default after controlling for the financial and demographic information. The L1 regression results corroborate the naïve Bayes findings (correlation between the two analyses is 0.582, $p < 0.01$. See details in Table A5). (3) To look beyond specific bi-grams and into the topics discussed in each loan we use a latent Dirichlet allocation (LDA) analysis. (4) Finally, relying on a well-known dictionary, the Linguistic Inquiry and Word Count (LIWC; Tausczik and Pennebaker 2010), we 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 most discriminate between loan requests by borrowers who default versus repay the loan in full, we ran a multinomial naïve Bayes classifier using the Python

scikit-learn 3.0 package on bi-grams (all possible words and bi-grams) that appeared in at least 400 loans (1,052 bi-grams). The classifier uses Bayes rule and the assumption of independence among words to estimate each word's likelihood of appearing in defaulted and paid loans. We then calculate the most "informative" bi-grams in terms of discriminating between defaulted and repaid loans by calculating the bi-grams with the highest ratio of P(bi-gram|defaulted)/P(bi-gram|repaid) and the highest ratio of P(bi-gram|repaid)/P(bi-gram|defaulted). Figures 2 and 3 present word clouds of the naïve Bayes analysis of bi-grams in their stemmed form (see underlying data in Tables A6a-b). The size of each bi-gram corresponds to the likelihood that it will be included in a repaid loan request versus a defaulted loan request (Figures 2), or in a defaulted loan request versus a repaid loan request (Figure 3). For example, the word "reinvest" in Figure 2 is 4.8 times more likely to appear in a fully paid than a defaulted loan request, while the word "god" in Figure 3 is 2.0 times more likely to appear in a defaulted than a fully paid loan request. The central cloud in each figure presents the most discriminant bi-grams (cutoff ratio = 1.5) and the satellite clouds represent emerging themes based on our grouping of these words.

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

We find that 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 "reinvest," "interest," and "tax;" (ii) Words that may be a sign of projected improvement in financial ability: "graduate," "wedding," and "promote;" (iii) Relative words such as "side," "rather," and "more than;" (iv) Long-term time related words such as "future," "every month," and "few years;" (v) "I" words such as "I'd," "I'll," "I'm." The above indicates that borrowers who paid in full may have nothing to hide, have a brighter future ahead of them, and are generally truthful. The latter insight is based on research showing the use of relative and time words as well as first person "I"

words is associated with greater candor because honest stories are usually more complex and personal (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). Borrowers who repaid their loan also used words that indicate their financial literacy (e.g., "reinvest," "after tax," and "minimum payment"). Indeed, higher financial literacy has been associated with lower debt (Brown et al. 2015).

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 "refinance,") and general hardship ("stress," "divorce," and "very hard"). 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 why,") and discussing their work state ("hard work," "worker"). Providing explanations is often connected to past deviant behavior (Sonenshein, Herzenstein, and Dholakia 2011). (iii) Appreciative and good-manner words toward lenders ("god bless," "hello") and pleading lenders for help ("need help," "please help"). Why is polite language more likely to appear in defaulted loan requests? One possibility is that it is not authentic. Indeed, Feldman et al. (2017) show that rude people are more trustworthy because their reactions seem more authentic. (iv) Referring to others such as "god," "son," "someone." 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). Further, reminders of god have been shown to increase risky behavior (Kupor, Laurin, and Levav 2015), alluding to the possibility these borrowers took a loan they were unable to repay. (v) Time related words ("total monthly," "day,") and future tense words

("would use," "will able"). 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 Lynch et al. (2010), who showed that long-term planning (versus short-term) is associated with lower procrastination, higher degree of assignment completion, and better credit scores. The mention of shorter horizon time words by defaulters is consistent with Shah, Mullainathan, and Shafir (2012) who find that financial resource scarcity leads people to shift their attention to the near future, neglect the distant future, and thus over-borrow. In sum, defaulting borrowers attempted to garner empathy and seem forthcoming and appreciative, but when it was time to repay their loan, they were unable to escape their reality. Interestingly, this narrative is very similar to the "Nigerian email scam", as described on the Federal Trade Commission's website: "Nigerian email scams are characterized by convincing sob stories, unfailingly polite language, and promises of a big payoff."[13]

While the naïve Bayes analysis is informative with respect to identifying words that are *associated* with loan default, for a practical use, we may wish to uncover the words and financial variables that are most *predictive* of default. To that end, we analyzed the variables with the highest importance in predicting repayment based on the Random Forest model used in the ensemble learning model (see Table A7). Most predictive of default are the financial variables such as lender rate and credit score, but also words such as "payday loan," "invest," "hard," "thank you," "explain," and "student." This is an interesting result because it implies that talking about "payday loans" is not fully redundant with the information provided by one's credit grade.

Relationship between Words Associated with Loan Default and Words Associated with Loan Funding

It is possible that some borrowers strategically use words to convince lenders to fund their

[13] We thank Gil Appel for this neat observation.

loans, therefore we examine whether lenders are aware of such strategic behavior, and if not, which words sipped through lenders' defenses and got them to fund overly-risky loans that eventually defaulted. To investigate the relationship between words associated with loan default and words related to loan funding, we ran a naïve Bayes analysis on the entire set of loans requests (122,479 funded and unfunded loan requests that included text), and assessed the bi-grams with the highest ratio of P(bi-gram|funded)/P(bi-gram|unfunded) and the highest ratio of P(bi-gram|unfunded)/P(bi-gram|funded). See Table A8 for summary statistics of this dataset and Table A9 for the naïve Bayes analysis results.

Figure 4 depicts for each bi-gram its value on the ratio P(bi-gram|defaulted)/P(bi-gram|repaid) versus its value on the ratio P(bi-gram|unfunded)/P(bi-gram|funded). We named a few representative bi-grams in the figure. A high correlation between the two ratios (P(bi-gram|defaulted)/P(bi-gram|repaid) and P(bi-gram|unfunded)/P(bi-gram|funded) means that lenders are largely aware of the words that are associated with default. Results show a fairly strong correlation between the two ratios ($r = 0.354$, $p < 0.01$), suggesting that lenders are at least somewhat rational when interpreting and incorporating the text in their funding decisions. Examining the results more carefully, we find roughly three types of words: (1) words that have high likelihood of default and low likelihood of funding (e.g., "need help," and "lost") or low likelihood of default and high likelihood of funding (e.g., "excellent credit," and "never miss") (2) words that have a stronger impact on loan funding than on loan default (e.g., words related to explanation, such as "loan explain," "situation explain," "explain what," and "explain why") and (3) words that "tricked" lenders to fund loans that were eventually defaulted (words that were related to default more than to loan funding). The most typical words in this category are "god" and "god bless" but also financial hardship words such as "payday," and "payday loan".

*** Insert Figure 4 around here ***

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

In Figures 2 and 3 we grouped the bi-grams into topics based on our own interpretation and judgment. However, several machine learning methods have been proposed to statistically combine words into topics based their common co-occurrence in documents. Probably the most commonly used topic modeling approach is the latent Dirichlet allocation analysis (LDA), which we apply on the complete dataset of all loan requests (the 122,479 funded and rejected loans).

We use the online variational inference algorithm for the LDA training (Hoffman, Bach, and Blei 2010), following Griffiths and Steyvers (2004)'s settings and priors. 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 (the perplexity measure), to determine the final number of topics. We find that the model with seven topics had the best fit (lowest perplexity). Table 3 presents the seven topics and the most representative words for each topic based on the relevance score of *0.5* (Sievert and Shirley 2014), Table A10 lists the top 30 words for each topic, and Figure A2 presents the perplexity analysis.

The topics we identify relate to the reason for the loan request, life circumstances, or writing style. We find three loan purpose topics: Employment and School, Interest Rate Consolidation, and Business and Real Estate. The other four topics are related to life circumstances and writing style: Expenses Explanation, Family, Loan Details Explanation, and Monthly Expenses. The monthly expenses topics are most likely related to a set of expenses Prosper recommended borrowers mention as part of their loan request during our data period.

Another advantage of the LDA is that it serves as a data-reduction technique, allowing us to use standard econometric methods (binary logistic regression) to relate text (topics) to default probabilities. Specifically, we ran a binary logistic regression with loan repayment = 1 and default = 0 as the dependent variable, the LDA probability of each topic appearing in the loan (the topic loan details explanation serves as benchmark), and the same set of textual information metrics, financial, and demographic variables used in the ensemble learning model described earlier. Table 4, presents the results of a binary logistic regression with LDA topics and controls. We first observe that the financial and demographic variables are significant and in the expected direction: repayment likelihood is increasing as credit grades improve, but decreasing with debt to income ratio, home ownership, loan amount, and lender rate. The strong relationship between lender rate and repayment suggests some level of efficiency among Prosper lenders.

*** Insert Tables 3, 4 around here ***

Relative to the topic Loan Detail Explanation, we find that topics of Employment and School, Interest Rate Reduction, and Monthly Payment are more likely to appear in repaid loan requests. These results corroborate the Naïve Bayes results that discussion of education is associated with lower default likelihood. Indeed, it is possible that traditional credit scores measures do not appropriately account for the positive effect of education on financial stability. The textual information provides lenders an indirect window into borrowers' educations. The Family topic, on the other hand, was less likely to appear in repaid loans. Consistent with the naïve Bayes analysis we find that the topic of explaining one's financials and loan motives, and referring to family 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). Finally, although not the purpose of the LDA analysis, we find that the binary logistic model that includes the seven LDA topics probabilities in addition to the financial information fits the data and predicts default better than a model that does not include the LDA results, though worse than the ensemble model (see Web Appendix for details).

In sum, multiple methods converge to uncover themes and words that differentiate defaulted from paid loan requests. Next, we explore whether borrowers' *writing styles* can shed light on the traits and states of defaulted borrowers.

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. 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). The result of almost two decades of research is lists of word categories that represent the writing style of people with different personalities (Kosinski, Stillwell, and Graepel 2013; Pennebaker and King 1999; Schwartz et al. 2013; Yarkoni 2010), mental health states (Preotiuc-Pietro et al. 2015), emotional states (Pennebaker, Mayne, and Francis 1997), as well as many other traits and states.

LIWC is composed of sub-dictionaries that can overlap (the same word can appear in several sub-dictionaries). We first calculate the proportion of stemmed words in each loan request that belong to each of the 64 dictionaries.[14] We then estimated a binary logit model (load

[14] 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.

repaid = 1 and loan defaulted = 0) to relate the proportions of words in each loan that appear in each dictionary to whether the loan was repaid, controlling for all financial and demographic variables used in our previous analyses. Results are presented in Table 5.

*** Insert Table 5 around here ***

We begin by noting that all financial and demographic control variables are in the expected direction and are consistent with our LDA results. Fourteen LIWC dictionaries were significantly related to repayment behavior, and several of them mirror the naïve Bayes and LDA results. To interpret our findings we rely on previous research that leveraged the LIWC dictionary, which allow us to conclude that defaulted loan requests contain words that are associated with the writing styles 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. This result is similar to our naïve Bayes findings. Indeed, past research shows liars are more likely to use present and future tense words because they represent unconcluded situations (Pasupathi 2007); (2) higher use of motion words (e.g., “drive,” “go,” and “run,”) which have been associated with lower cognitive complexity (Newman et al. 2003), and lower use of relative words (e.g., “closer,” “higher,” and “older,”) which are associated with higher complexity (Pennebaker and King 1999). Deceptive language has been shown to include more motion words and fewer relative words because it is less complex in nature; (3) Similar to our finding from the naïve Bayes analysis that defaulters tend to refer to others and repayers tend to use more “I” words, 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 Bond and Lee 2005; Newman et al. 2003). We note that in the context of hotel reviews, Ott, Cardie, and Hancock (2012) find higher use of "I" in fake reviews possibly because they did not have much to write about the hotel itself (because they have never been there) so they described their own activities; (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 because the use of such words seem to draw attention away from the self (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). But, do borrowers intentionally lie to lenders? One possibility is that people can predict, with some accuracy, future default, months, or even years ahead. If so, this would support the idea that defaulters are writing either intentionally or unconsciously to deceive lenders. However, there is another option: borrowers on Prosper might genuinely believe that they will be able to pay the borrowed money in full. Indeed, past research suggests that people are often optimistic about future outcomes (Weinstein 1980). What they may be hiding from lenders is the extent of their difficult situations and circumstances.

Our second observation is that the sub-dictionaries associated with the writing style of extroverts are also associated with greater likelihood to default. Extroverts have been shown to use more religious and body related words (e.g., "mouth," "rib," "sweat," and "naked;" Yarkoni 2010), social and humans words (e.g., "adults," "boy," and "female;" Hirsh and Peterson 2009; Pennebaker and King 1999; Schwartz et al. 2013; Yarkoni 2010), motion words (e.g., "drive,"

"go," and "run;" Schwartz et al. 2013), achievement words (e.g., "able," "accomplish," and "master,") and fewer filler words (e.g., "blah" and "like;" Mairesse et al. 2007)—all of which are significantly related to a greater likelihood of default in our analysis (see Table 5).

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 (Nyhus and Webley 2001). Moreover, it is not a coincidence that the fourteen LIWC dictionaries that were significantly correlated with default are correlated with both extroversion and 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).

We did not find consistent and conclusive relationship between the LIWC dictionaries associated with each of the other big five 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. Finally, we acknowledge that there may be variables that are confounded with both the observable text and unobservable personality traits or states 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 and predicts default better than a model that does not include the textual information (see Web Appendix for details).

GENERAL DISCUSSION

The words we write matter. Aggregated text has been shown to predict market trends (Bollen, Mao, and Zeng 2011), market structure (Netzer et al. 2012), virility of news articles (Berger and Milkman 2012), prices of 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 writers (Ventrella 2011), identify liars (Newman et al. 2003) and fake reviews (Ott, Cardie, and Hancok 2012), assess personality traits (Schwartz et al. 2013; Yarkoni 2010), and the mental state (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 their predictive ability. Using machine learning methods we uncover the words and topics borrowers often include in their loan request. We find that at loan origination, defaulters used simple but wordier language, wrote about hardship, 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 the commonly used LIWC dictionary we infer that defaulting borrowers write similarly to people who are extroverts and to those who lie. 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 the riskiness of different types of loans and borrowers. Simply put, we show that borrowers, consciously or not, leave traces of their intentions, circumstances, and personality in the text they write when applying for a loan—a digital involuntary "sweat".

Theoretical and Practical Contribution

Our work contributes to the recent but growing marketing literature on uncovering behavioral insights on consumers from the text they write and the traces they leave on social media (Humphreys and Jen-Hui Wang 2018; Matz and Netzer 2017). We demonstrate that text

consumers write at loan origination is indicative of their states and traits, and predictive of their future repayment behavior. In an environment characterized by high uncertainty and high stakes, 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"—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 by showing that people with different economic and financial situations use words differently.

Second, we make an additional contribution to the text analytics literature. The 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) fake review (Ott, Cardie and Hancock. 2012), but rarely on predicting distant future behavior of writers.

Third, our approach to predicting default relies on an automatic algorithm that mines individual words, including those without much meaning (e.g., articles and fillers), in the entire textual corpora. Past work on narratives used to facilitate economic transaction (e.g., Herzenstein, Sonenshein, and Dholokia 2011; Martens, Jennings, and Jennings 2007), employed human coders and therefore is prone to human mistakes, is not scalable, which limits its predictive ability and practical use. It may not be surprising that people who describe hardship in the text at loan origination are more likely to default—as past research has shown. However, our work demonstrates that defaulters and repayers also differ in their usage of pronouns and tenses, and those seemingly harmless pronouns have the ability to predict future economic behaviors. Research shows that unless agents are very mindful regarding their usage of pronouns, it is

unlikely that they noticed (not to mention planned) to use specific pronouns in order to manipulate the reader (Pennebaker 2011)—lenders in our case.

Fourth, we provide evidence that our method of automatically analyzing free text is an effective way of supplementing traditional measures and even replacing some aspects of the human interaction of traditional bank loans. Textual information, such as the one we analyze, not only sheds light on past behavior (which may be captured by measures such as credit scores) and provides it with context, but also offers information about the future, which is unlikely to be captured by credit history reports. These future events may be positive (e.g., graduating, starting a new job) or negative (e.g., impending medical cost), and certainly affect lending decisions.

Furthermore, because lending institutions place a great deal of emphasis on the importance of models for credit risk measurement and management, they have 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. Such endeavors may be particularly useful in "thin file" situations where historical data about customers' finances is sparse. A GAO (Government Accountability Office) report from December 2018, written at the request of Congress, delineates "alternative data usage" in loan decisions by fintech lenders (Prosper.com is one of the 11 lenders surveyed)[15]. Alternative data is defined as anything not used by the three credit bureaus, and some examples range from rent payments, education and alma mater, to internet browser history and social media activity. These data are either given freely by potential borrowers or are bought from data aggregators. According to the report, currently there are no specific regulations regarding the usage of these data and fintech lenders use consulting and law firms specializing in

[15] https://www.gao.gov/assets/700/696149.pdf

fair lending issues to test their models for compliance with fair lending laws. Fintech lenders explain that alternative data allow them to lend money to those who otherwise would not be eligible, including small businesses requiring a small loan, and individuals with low credit scores. Hence, while fintech lenders voiced their need for more guidance from the CFPB, the use of alternative data is unlikely to go away, and if anything, their reliance on it will only increase.

The objective of the current research is to explore the predictive and informative value of the text in loan applications, however, as mentioned in the GAO report, using such information to decide which borrowers should be granted loans must be carefully assessed and comply with fair lending laws and their interpretation by the CFPB.

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 important 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), should be invariant to the type of loan and lending platform, whereas other aspects may vary by context. Yet, since we analyze data from one platform, we present a case study for the potential value of textual analysis in predicting default. Future research should extend our work to different populations, other types of unsecured loans (e.g., credit card debt), as well as secured loans, (e.g., mortgages).

Second, our results should be extended to other types of media and communication, such as phone calls or online chats. It would be interesting to test how an active conversation, two-sided correspondence versus one-sided input, may affect the results. In our data, the content of

the text is entirely up to borrowers—they disclose what they wish in whichever writing style they choose. In a conversation, the borrower may be prompted to provide certain information.

Third, we document specific words 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. In other words, the Lucas critique (Lucas 1976) that historical data cannot predict a change in economic policy, because future behavior changes as policy changes, may apply to our situation. In what follows we discuss this critique on both demand for loans (the borrowers' side) and supply of loans (lenders' side). We begin with borrowers and wish to emphasize two aspects. First, an important objective of our study is to explore and uncover the behavioral signals or traces in the text that are associated with loan default. This objective is descriptive rather than prescriptive and hence not susceptible to the Lucas critique. Second, for the Lucas critique to manifest three conditions need to occur: (1) the research findings need to fully disseminate, (2) the agent needs to have sufficient incentive to change her behavior, and (3) the agent needs to be able to change her behavior based on the proposed findings (Van Heerde, Dekimpe, and Putsis 2005). With respect to the first point, evidence from deception detection mechanisms (Li et al. 2014) as well as macroeconomics (Rudebusch 2005) suggest that dissemination of research results rarely fully affect behavior. Moreover, borrowers in our context vary substantially in terms their socio-economic and education level, which makes it nearly impossible to change the behavior of all or even the majority of them. As for the second and third points, borrowers indeed have a strong incentive to change their behavior due to the decision importance, however, they might not necessarily be able to do that. Even if our results are fully disseminated, people use of pronouns and tenses (the

finding at the heart of our research) has been found to be largely subconscious and changing it demands heavy cognitive load (Pennabaker 2011). Considering the supply of loans, we acknowledge that lenders have an incentive to learn about possible tactics borrowers employ in order to garner trust and empathy when lending is economically not warranted. If lenders have the ability to learn about word usage that enables bad loan requests to be sufficiently similar to good ones and hence get funded (when they should not have been), they are likely fund a different mix of loans. Put differently, if our findings are disseminated then a different set of loans would have been funded—some that were rejected would have been funded and some that were funded would have been rejected. Future research should explore how word usage changes over time and the resulting mix of funded and unfunded loans.

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

To conclude, borrowers leave meaningful signals in the text of loan applications that help predict default, sometimes years post loan origination. Our research adds to the literature utilizing text mining to better understand consumer behavior (Humphreys and Jen-Hui Wang 2018), and especially in the realm of consumer finance (Herzenstein, Sonenshein, and Dholakia 2011).

Epilogue: We thank you and bless you for reading our paper, we politely ask that you cite our work, and we promise to work hard and cite you back.

REFERENCES

Agarwal, Sumit and Robert Hauswald (2010), "Distance and Private Information in Lending," *Review of Financial Studies*, 23(7), 2757-2788.

Agarwal, Sumit, Paige M. Skiba, and Jeremy Tobacman (2009), "Payday Loans and Credit Cards: New liquidity and Credit Scoring Puzzles" *American Economic Review*, 99(2), 412-17.

Anderson, Jon, Stephen Burks, Colin DeYoung, and Aldo Rustichini (2011), "Toward the Integration of Personality Theory and Decision Theory in the Explanation of Economic Behavior." Presented at the *IZA workshop: Cognitive and non-cognitive skills*.

Arya, Shweta, Catherine Eckel, and Colin Wichman (2013), "Anatomy of the Credit Score," *Journal of Economic Behavior and Organization* 95, 175-185.

Avery, Robert B., Raphael W. Bostic, Paul S. Calem, and Glenn B. Canner (2000), "Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files," *Real Estate Economics*, 28(3), 523-547.

Banko, Michele and Eric Brill (2001), "Scaling to Very Very Large Corpora for Natural Language Disambiguation," In *Proceedings of the 39th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 26-33.

Barber, Brad M. and Terrance Odean (2001), "Boys will be Boys: Gender, Overconfidence, and Common Stock Investment," *The Quarterly Journal of Economics*, 116 (1), 261-292.

Berger, Jonah, and Katherine L. Milkman (2012), "What Makes Online Content Viral?" *Journal of Marketing Research*, 49 (2), 192-205.

Bergstra, James and Yoshua Bengio (2012), "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, 13 (Feb), 281-305.

Berneth, Jeremy, Shannon G. Taylor, and Harvell Jackson Walker (2011), "From Measure to Construct: An Investigation of the Nomological Network of Credit Scores" *In Academy of Management Proceedings*, 1-6.

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, 313-29.

Brown, Meta, John Grigsby, Wilbert van der Klaauw, Jaya Wen, and Basit Zafar (2015), "Financial Education and the Debt Behavior of the Young," *Federal Reserve Bank (New York)*, Report #634.

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.

Dholakia, Utpal, Leona Tam, Sunyee Yoon, and Nancy Wong (2016), "The Ant and the Grasshopper: Understanding Personal Saving Orientation of Consumers," *Journal of Consumer Research,* 43 (1), 134-155.

Farrell, Joseph and Matthew Rabin (1996), "Cheap Talk," *The Journal of Economic Perspectives*, 10 (3), 103-118.

Feldman, Gilad, Huiwen Lian, Michal Kosinski, and David Stillwell (2017), "Frankly, We Do Give A Damn: The Relationship Between Profanity and Honesty," *Social Psychological and Personality Science*, 8 (7), 816-826.

Fernandes, Daniel, John G. Lynch Jr, and Richard G. Netemeyer (2014), "Financial Literacy, Financial Education, and Downstream Financial Behaviors," *Management Science*, 60(8), 1861-83.

Freitas, Antonio L., Nira Liberman, Peter Salovey, and E. Tory Higgins (2002), "When to Begin? Regulatory Focus and Initiating Goal Pursuit," *Personality and Social Psychology Bulletin,* 28(1), 121-130.

Griffiths, Thomas L. and Mark Steyvers (2004), "Finding Scientific Topics," *Proceedings of the National academy of Sciences*, 101 (1), 5228-5235.

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.

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, 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*, 48(SPL), S138-S149.

Hildebrand, Thomas, Manju Puri, and Jörg Rocholl (2017), "Adverse Incentives in Crowdfunding," Management Science, 63 (3), 587-608.

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.

Humphreys, Ashlee and Rebecca Jen-Hui Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 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).

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), 374-84.

Lee, Dokyun, Kartik Hosanagar, and Harikesh S. Nair (2018), "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," *Management Science*, 64 (11), 5105-31

Li, Jiwei, Myle Ott, Claire Cardie, and Eduard Hovy (2014), "Towards a General Rule for Identifying

Deceptive Opinion Spam," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1566-1576.

Lucas, Robert (1976), "Econometric Policy Evaluation: A Critique," *The Phillips Curve and Labor Markets, Carnegie-Rochester Conference Series on Public Policy*. New York: Elsevier, 19–46.

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.

Matz, Sandra and Oded Netzer (2017), "Using Big Data as a Window into Consumers' Psychology," *Current Opinion in Behavioral Sciences*, 18 (December), 7-12,

Mayer, Christopher, Karen Pence, and Shane M. Sherlund (2009), "The Rise in Mortgage Defaults," *The Journal of Economic Perspectives*, 23 (1), 27-50.

McLaughlin, G. Harry (1969), "SMOG Grading—A New Readability Formula," *Journal of Reading*, 12(8), 639-646.

McAdams, Dan P. (2001), "The Psychology of Life Stories," *Review of General Psychology*, 5(2), 100-123.

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.

Netemeyer, Richard G., Dee Warmath, Daniel Fernandes, and John Lynch Jr. (2018), "How Am I Doing? Perceived Financial Well-Being, Its Potential Antecedents, and Its Relation to Overall Well-Being," *Journal of Consumer Research*, Forthcoming.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance through Text Mining," *Marketing Science,* 31(3), 521-43.

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.

Norvilitis, Jill M., Michelle M. Merwin, Timothy M. Osberg, Patricia V. Roehling, Paul Young, and Michele M. Kamas (2006), "Personality Factors, Money Attitudes, Financial Knowledge, and Credit-card Debt in College Students," *Journal of Applied Social Psychology*, 36(6), 1395.

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," *Proceedings of the 21st Conference on World Wide Web*, 201-210.

Palmer, Christopher (2015), "Why did so many subprime borrowers default during the crisis: Loose credit or plummeting prices?" *Working Paper*, University of California Berkeley.

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. (2011), "The Secret Life of Pronouns," *New Scientist*, 211(2828), 42-45.

Pennebaker, James W. and Lori D. Stone (2003), "Words of Wisdom: Language Use over the Life Span," *Journal of Personality and Social Psycholo*gy, 85 (2), 291-301.

Pennebaker, James W. and Anna Graybeal (2001), "Patterns of Natural Language Use: Disclosure, Personality, and Social Integration," *Current Directions in Psychological Science*, 10 (3), 90-3.

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.

Pompian, Michael M. and John M. Longo (2004), "A New Paradigm for Practical Application of Behavioral Finance: Creating Investment Programs Based on Personality Type and Gender to Produce Better Investment Outcomes," *Journal of Wealth Management*, 7, 9-15.

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, et al. (2015), "The Role of Personality, Age and Gender in Tweeting about Mental Illnesses," *NAACL HLT*, 21-30.

Rudebusch, Glenn D. (2005), "Assessing the Lucas Critique in Monetary Policy Models." *Journal of Money, Credit, and Banking*, 37(2), 245-272.

Rugh, Jacob S., and Douglas S. Massey (2010), "Racial Segregation and the American Foreclosure Crisis," *American Sociological Review*, 75 (5), 629-651.

Rustichini Aldo, Colin DeYoung, Jon E. Anderson, Stephen Burks (2016), "Toward the Integration of Personality Theory and Decision Theory in Explaining Economic Behavior: An Experimental Investigation," *Journal of Behavioral and Experimental Economics*, 64, 122-137.

Schwartz, H. Andrew, et al. (2013), "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *Plos One*, 8 (9), e73791.

Sengupta, Rajdeep, and Geetesh Bhardwaj (2015), "Credit Scoring and Loan Default." *International Review of Finance*, 15 (2), 139-167.

Shah, Anuj K., Sendhil Mullainathan, and Eldar Shafir (2012), "Some Consequences of Having Too Little," *Science*, 338 (6107), 682-685.

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.

Sonenshein, Scott, Michal Herzenstein, and Utpal M. Dholakia (2011), "How Accounts Shape Lending Decisions Through Fostering Perceived Trustworthiness," *Organizational Behavior and Human Decision Processes*, 115 (1), 69-84.

Sussman, Abigail B. and Rourke L. O'Brien (2016), "Knowing When to Spend: Unintended Financial Consequences of Earmarking to Encourage Savings," *Journal of Marketing Research,* 53 (5), 790-803.

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.

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.

Tumasjan, Andranik, et al. (2010), "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment," *ICWSM*, 178-185.

Van Heerde, Harald J., Marnik G. Dekimpe, and William P. Putsis Jr. (2005), "Marketing Models and the Lucas Critique," *Journal of Marketing Research*, 42 (1), 15-21.

Ventrella, Jeffrey J. (2011), *Virtual Body Language*, ETC Press.

Wei, Yanhao, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas (2015), "Credit Scoring with Social Network Data," *Marketing Science*, 35 (2), 234-258.

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.

Weinstein, Neil D. (1980) "Unrealistic Optimism about Future Life Events," *Journal of Personality and Social Psychology*, 39(5), 806-820.

Whalen, Sean and Gaurav Pandey (2013), "A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics," *In Data Mining (ICDM), 2013 IEEE Conference, 807-816.*

Yarkoni, Tal (2010), "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.

Table 1. Descriptive statistics for funded loan requests (n = 19,446)

Variables	Min	Max	Mean	SD	Freq.
Amount requested	1,000	25,000	6,506.9	5,732.4	
Debt-to-income ratio	0	10.01	.33	.86	
Lender interest rate	0	.350	.180	.077	
Number of words in description	0	766	207.9	137.4	
Number of words in title	0	13	4.595	2.012	
% of long words (6+ letters)	0%	71.4%	29.8%	6.4%	
SMOG	3.129	12	11.347	1.045	
Enchant spellchecker	0	54	2.355	3.015	
# Prior listings	0	67	2.016	3.097	
Credit grade: AA					0.131
A					0.134
B					0.174
C					0.215
D					0.182
E					0.084
HR					0.081
Loan repayment (1 = paid, 0 = defaulted)					0.669
Loan image dummy					0.666
Debt-to-income missing dummy					0.05
Group Membership Dummy					0.253
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/ demo	(3) Text & financial/demo	Improvement from (2) to (3)
Low credit grades: E, HR	61.78%	62.62%	65.61%	4.77%*
Medium credit grades: B, C, D	62.84%	65.75%	68.06%	3.51%*
High credit grades: AA, A	72.34%	77.38%	78.88%	1.94%*
Overall AUC	66.69%	70.72%	72.60%	2.64%*
Jaccard Index	37.97%	37.01%	39.50%	3.92%*
AUC of the underlying models of the ensemble				
Logistic L1	67.75%	70.09%	71.66%	2.23%*
Logistic L2	68.09%	68.58%	72.09%	5.10%*
Random Forest (Variance Selection)	64.62%	70.35%	70.85%	0.71%*
Random Forest (Best Features Selection)	66.14%	69.24%	71.13%	2.73%*
Extremely Randomized Trees (Extra Trees)	66.65%	69.81%	70.98%	1.66%*

Notes: all AUCs 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 on randomly selected 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 P-value < 0.01 level.

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

LDA topic	Words with highest relevance ($\lambda = 0.5$)
Employment and School	Work, Job, Full, School, Year, College, Income, Employ, Student
Interest Rate Reduction	Debt, Interest, Rate, High, Consolidate, Score, Improve, Lower
Expenses Explanation	Expense, Explain, Cloth, Entertainment, Cable, Why, Utility, Insurance, Monthly
Business and Real Estate	Business, Purchase, Company, Invest, Fund, Addition, Property, Market, Build, Cost, Sell
Family	Bill, Try, Family, Life, Husband, Medical, Reality, Care, Give, Children, Hard, Daughter, Chance, Son, Money, Divorce
Loan Details and Explanations	Loan, Because, Candidate, Situation, Financial, Purpose, House, Expense, Monthly, Income
Monthly Payment	Month, Payment, Paid, Total, Account, Rent, Mortgage, Save, List, Every, Payday, Budget

Note: the sample words are chosen based on the relevance measure with $\lambda = 0.5$. See Table A10 for more exhaustive lists of words for each topic.

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

Financial and loan related variables	Estimate (Std. E)	Textual variables	Estimate (Std. E)
Amount Requested (in 10^5)	**-7.08 (0.35)**	Number of words in Description (in 10^4)	-0.00 (0.001)
Credit Grade HR	**-0.79 (0.08)**	Number of spelling mistakes	0.00 (0.00)
Credit Grade E	**-0.43 (0.08)**	SMOG (in 10^3)	-1.8 1.6
Credit Grade D	**-0.33 (0.06)**	**Words with 6 letters or more**	**-0.72 (0.37)**
Credit Grade C	**-0.17 (0.05)**	Number of words in the title (in 10^3)	-7.70 (8.00)
Credit Grade A	**0.76 (0.08)**	**Employment and school**	**2.26 (0.43)**
Credit Grade AA	**0.24 (0.07)**	**Interest rate reduction**	**2.87 (0.43)**
Debt to income	**-0.09 (0.02)**	Expenses explanation	-0.25 (0.60)
Images	0.04 (0.04)	Business and real estate loan	0.64 (0.39)
Home owner status	**-0.36 (0.04)**	**Family**	**-1.16 (0.43)**
Lender interest rate	**-5.57 (0.30)**	Monthly payments	0.99 (0.41)
Bank draft fee annual rate	**-39.12 (18.5)**		
Debt to income missing	**-0.25 (0.07)**		
Group membership	**-0.20 (0.04)**		
Prior listings	**-0.03 (0.01)**	**Intercept**	**2.13 (0.35)**

* Bold face for P-value ≤ 0.05. S.E in parenthesis. For brevity we do not report estimates for location, age, gender, and race.

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

Financial and basic text variables:		**LIWC dictionaries**					
Amount Requested(x 10^5)	**-6.776 (0.352)**	Swear words	35.3585 (34.838)	Past words	-1.4975 (1.923)	Person pronoun words	-0.1517 (6.466)
Credit Grade HR	**-0.8030 (0.082)**	**Filler words**	**13.5714 (6.000)**	Inhibition words	-4.0838 (3.292)	Work words	0.5035 (0.898)
Credit Grade E	**-0.4177 (0.080)**	Perception words	14.2621 (10.174)	Home words	-2.964 (1.693)	Sexual words	-10.6703 (10.580)
Credit Grade D	**-0.3088 (0.061)**	**Relative words**	**9.6016 (2.240)**	Hear words	-4.1232 (13.237)	They words	-14.4516 (9.076)
Credit Grade C	**-0.15879 (0.054)**	Friend words	7.5905 (6.698)	I words	-1.368 (7.976)	Positive emotion words	0.4510 (1.958)
Credit Grade A	**0.7489 (0.076)**	Anxiety words	10.9341 (8.662)	Tentative words	-3.0452 (1.968)	Money words	0.4857 (0.720)
Credit Grade AA	**0.2679 (0.066)**	**Negate words**	**7.3474 (3.215)**	Non-fluency words	-3.438 (9.242))	Ingest words	-3.8567 (5.029)
Debt to income	**-0.08426 (0.010)**	Insight words	3.1410 (2.712)	Anger words	-0.8509 (9.514)	Verbs words	0.1246 (1.260)
Debt to income Missing	**-0.2391 (0.074)**	We words	4.7593 (8.153)	Achieve words	-2.747 (1.510)	Adverbs words	-0.3043 (1.823)
Images	0.0468 (0.038)	Pronoun words	3.5098 (9.758)	Inclusion words	-4.1367 (2.236)	Functional words	-1.1873 (1.791)
Home owner status	**-0.3493 (0.037)**	Exclusion words	3.4075 (2.662)	She/he words	-3.3432 (7.128)	Bios words	-1.6958 (2.576)
Lender interest rate	**-5.6184 (0.301)**	Sad words	-0.11855 (6.276)	You words	-1.0086 (8.566)	Assent words	-5.8254 (13.831)
Bank draft fee annual Rate	**-40.4880 (18.623)**	Quantitative words	2.5479 (1.870)	Cause words	-3.7248 (2.407)	Family words	-1.8106 (2.559)
Prior listings	0.0008 (0.012)	Articles	3.8222 (2.018)	**Social words**	**-4.133 (1.503)**	I pronoun words	-1.5238 (9.783)
Group membership	**-0.1627 (0.042)**	Numbers words	2.3621 (2.640)	Health words	-4.3656 (4.140)	Death words	-18.7405 (10.460)
Number of words in Description	-0.0015 (0.001)	Preposition words	2.2468 (1.779)	**Certain words**	**-5.3928 (2.621)**	**Body words**	**-16.4249 (5.637)**
Number of spelling mistakes	**-0.0131 (0.006)**	Conjoint words	2.0789 (1.779)	**Present words**	**-5.3329 (1.625)**	**Religion words**	**-19.8774 (6.590)**
SMOG	0.0343 (0.018)	Auxiliary verbs words	0.8394 (2.289)	**Human words**	**-7.1079 (3.449)**	**Feel words**	**-27.2482 (11.078)**
Words with 6 letters or more	-0.6324 (0.508)	Affect words	1.7009 (1.743)	**Space words**	**-8.2879 (2.440)**	See words	-12.2565 (10.94)
Number of words in the title	-0.0070 (0.008)	Discrepancy words	0.9167 (2.583)	**Future words**	**-7.1254 (3.408)**	Leisure words	-1.18358 (2.420)
		Cognitive mechanism words	1.277 (1.795)	**Motion words**	**-9.4886 (2.688)**		
Intercept	**3.0565 (0.435)**	Negative emotion words	0.8531 (4.601)	**Time words**	**-9.4991 (2.193)**		

* Bold face for P ≤ 0.05. S.E in Parentheses. For brevity we do not report estimates for 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

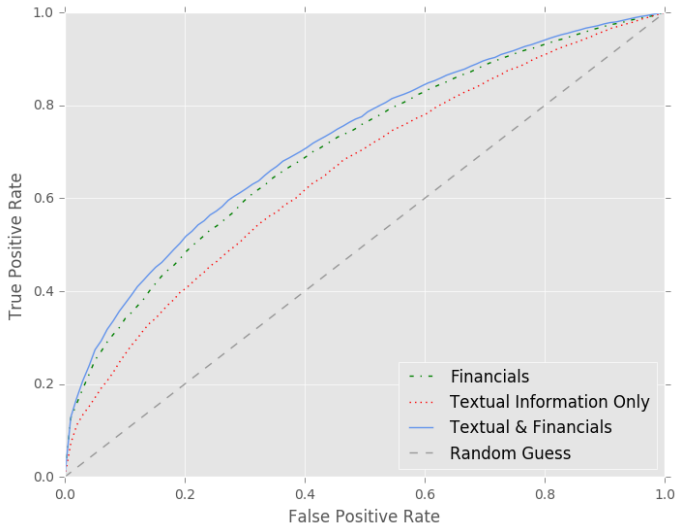
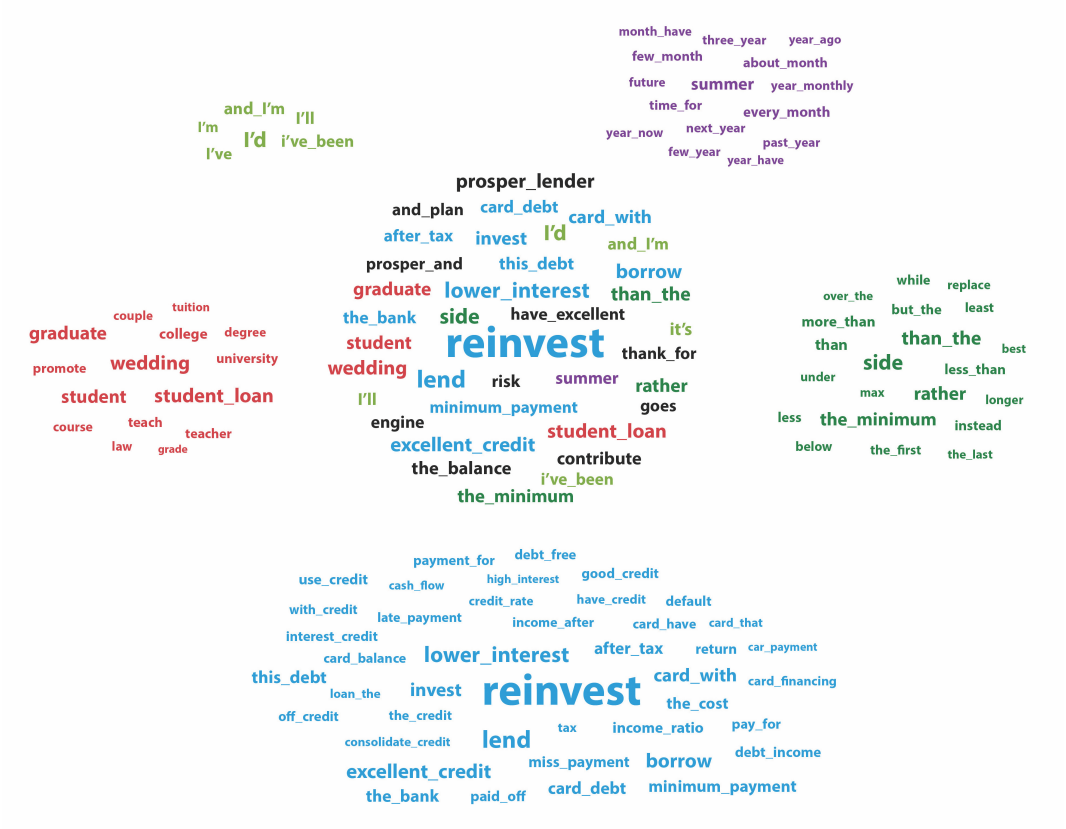
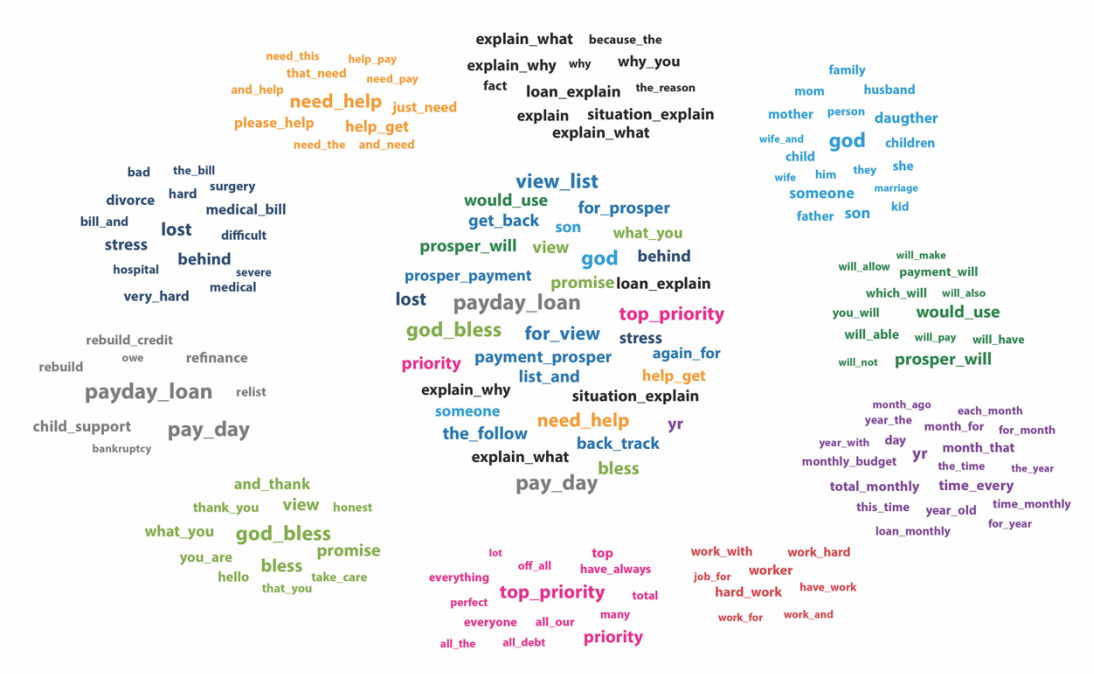


Figure 2. Words indicative of loan repayment



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

Figure 3. Words indicative of loan default



Note: The most common words appear in the middle cloud (cutoff = 1:1.5) and then organized by themes. On the top, in black, and clockwise: words related to explanations, external influence and others, future tense, time, work, extremity, appealing to lenders, financial hardship, hardship, and desperation and plea.

Figure 4: Naïve Bayes analysis for loan funding and loan default