Vocabulary as an indicator of creditworthiness: An analysis of public loan data

Justin Wagers

University of Puget Sound, Department of Economics

***Abstract*— The purpose of this research is to determine the usefulness of a borrower's vocabulary in determining his/her creditworthiness. The analysis takes a word-frequency approach to 36,055 loans from the peer-to-peer lending platform Lending Club, and evaluates text submitted by borrowers to improve the prediction of whether they will pay back their loan through a nave-Bayes classifier model. Vocabulary, when paired with traditional creditworthiness measures, is found to significantly improve the prediction accuracy of a borrower's creditworthiness as compared to the accuracy of traditional credit measures alone.**

I. Introduction

The common credit score has become a default indicator when it comes to determining how trustworthy a borrower might be. The financial industry is quite dependent on the credit score: banks, credit card companies, and mortgage lenders all use it as a major tool when determining whether a customer will receive a loan. However, some evidence shows that the credit score might not be the precise tool that it is made out to be; credit scoring has had particular accuracy problems with underrepresented groups. The industry is always looking for tools to improve knowledge about a borrower before lending to them; there is a continual search for ways in which to improve the imperfect credit score. This search for improvement has led to the very recent analysis of a borrower's vocabulary as a potential indicator of their creditworthiness. Often borrowers submit a description of their purpose when applying for a loan, but this text has hardly been utilized by lending entities. However, scholarly literature in psychology suggests that there is power in using vocabulary as a predictor of future behavior. Thus, the vocabulary of borrowers, in the form of a typed loan description, might provide additional accuracy to the credit score in estimating the probability of default for a given loan.

The peer-to-peer lending industry, by making all of its loan data publicly available, presents an opportunity to analyze the links between text submitted by borrowers and default probability. Investors on these sites have taken advantage of the publicly available loan data to improve the returns on their portfolios primarily by taking into account additional numerical information on borrowers (inquiries, delinquencies, income, etc.). But only recently have they begun to see the text descriptions as a useful indicator of creditworthiness. Peer-to-peer investors have just begun to dig for patterns in the loan descriptions, looking at simple characteristics such as description length. However, a much more in-depth analysis is needed to determine if there existed any relationships between the actual content of these descriptions and default rates of borrowers. Previous research utilizing non-linear regression established a correlation between individual word usage and default rate for a set of 17 words of high frequency in loan descriptions, some indicating a higher chance of default and some indicating a lower chance (Wagers 2016). So, given this evidence of correlation for individual words, the next question is whether it is possible to use these words collectively to improve the current estimate of default probability as determined by credit score.

An attempt to answer this question requires a predictive model that can smoothly incorporate natural language in conjunction with numerical variables. A Bayesian statistical approach offers this ability, and is commonly used in the field of natural language analysis in general. Bayesian statistics, as opposed to the more commonly used frequentist statistics, is advantageous for its ability to incorporate a prior belief in the construction of an accurate predictive model. I use a nave Bayes classifier predictive model, perhaps the most commonly used model for natural language analysis, to assess the ability to improve a prediction of default probability given a borrower's loan description. This link between vocabulary and creditworthiness would be exciting because it could support the importance of non-numerical variables for financial prediction purposes: by constructing a model that quantifies previously overlooked data, financial predictions can become more precise. In particular, lenders could use patterns in vocabulary to improve knowledge on lenders beyond the numerical credit score.

II. Background on Peer-to-Peer Lending

The dataset used in this study is publicly available data from the Lending Club platform. Lending Club is the biggest peer-to-peer lending platform today, with nearly $16 billion in loans funded since their start in 2007 (Lending Club Statistics, 2015). Lending Club and other peer-to-peer lending platforms have eliminated the need for a bank as a middle man, providing individual borrowers with lower interest rates and individual lenders higher returns than they would receive by parking their money in a savings account (Athwal, 2014). Borrowers on Lending Club are essentially crowd-funded by investors in $25 increments, and investors in turn receive interest from the borrowers. Lending Club evaluates a borrower's credit information including debt-to-income ratio, number of bankruptcies, and credit score. These variables give Lending Club some insight into a

borrower's trustworthiness, which allows them to assign an interest rate to each loan. In addition to this information, borrowers submit a short description of why they are taking out a loan; these descriptions will serve as the substance for this analysis.

III. Literature Review

This area of research is especially important given that both peer-to-peer lending and natural language analysis are on the rise. The US peer-to-peer lending market generated $6.6 billion in loans in 2015, up an enormous 128% from 2014 (Bakkler 2016). As the market grows, so will the pool of investors, and these investors will be looking for new strategies to take advantage of the large quantities of data available on these platforms. After they have evaluated all of the numerical characteristics of borrowers, some will turn (and already have turned) to other ways to increase their returns; analysis of natural language might provide the next best way to infer information about a borrower. To understand how natural language could improve the estimation of creditworthiness on peer-to-peer sites, it is important to evaluate the credit score and its effectiveness, how verbal and written language analysis has been conducted in psychology and economics, past investing strategies on peer-to-peer lending, and the advantages of a Bayesian statistical model with natural language.

The Fair Isaac Corporation started working on a standardized measure of creditworthiness in 1954, and began use of the FICO score in 1989; a remarkably similar formula is now used by all three major credit-reporting agencies in the U.S. (Trainor, 2015). The systematic reliance on the FICO score cannot be understated: credit score helps determine insurance rates, employment options, protection against fraud, and the ability to borrow money in any form for consumers (Hamm, 2014). From a lender's standpoint, credit score is the critical measure of creditworthiness; many have a concrete credit score threshold to determine who can borrow certain amounts of money. However, credit score is not a 100% accurate reflection of true creditworthiness. Mester (1997) emphasizes that there are some cases where people high credit scores will default, and vice versa. Mester also points out a number of flaws with credit scoring, in particular how it inaccurately estimates the creditworthiness of underrepresented groups (Mester 1997). Borrowers in the peer-to-peer lending industry aren't forced to rely on this sometimes flawed measure of creditworthiness; the industry has differentiated itself in that all loan data has been made available to the public, allowing investors to analyze information about the borrowers themselves to determine what characteristics lead to a trustworthy borrower. Among this information is a description submitted by borrowers. Thus, an opportunity has presented itself to quantify the links between natural language, in the form of loan descriptions, and creditworthiness, in the form of default rates on these loans.

The study of language and its effect on behavior are frequently discussed in social psychology. Many of the original theories linking language and behavior are credited to Kenneth Pike's 1954 Language in Relation to a Unified Theory of the Structure of Human Behavior, Pike theorized that language and behavior were too ingrained within each other to consider them separately; The activity of man constitutes a structural whole with language in a behavioral compartment insulated in character, content, and organization from other behavior (Pike 1954). This theory that combines language and behavior into one should alone be reason enough to investigate the prediction power of language. However, while Pike was a forefront theorist on language and behavior, he left the door open for more empirical work on the interworkings between the two. Researchers have continually kept Pike's theory in mind while investigating more concrete relationships within language analysis. Lera Boroditsky, a professor of Cognitive Science at UCSD, has provided evidence in various studies for the effect of language on visual perception, risk taking and the way people perceive events. Most notably, Boroditsky has thoroughly investigated correlations between language and time perception. She provides strong evidence that speakers of different languages perceive time in different terms: English speakers think in length of time, while Mandarin speakers think of time vertically, which can influence whether and how speakers of each language plan for the future (Boroditsky 2009). Her observations are notable because they indicate the ability to correlate language with future behavior. This finding is further supported by economist Keith Chen's determination that language can be an indicator of economic behavior, namely savings rates of individuals. Chen found that speakers of languages with obligatory future-time reference had higher savings rates than speakers of languages that didn't force them to reflect on time scale when speaking (Chen 2013). Chen's results are especially important because they suggest that language can in some ways be a predictor of economic behavior, which is particularly relevant to this study. However, both Boroditsky and Chen focus strictly on verbal language, while I am primarily concerned about the written language that borrowers submit in their loan applications.

Luckily, written language analysis has shown similar significance. One 2010 study showed that written word choice has proven useful in determining personal information about individuals, including one's opinions on controversial issues. Klebanov et. al. (2010) conducted a word frequency analysis of opinion-based text from abortion debates, death penalty blogs, and film reviews on Bitter Lemons to investigate the extent to which vocabulary was not only a matter of topic but was reflective of an individual's perspective on an issue. The researchers found that use of a small number of keywords could signal an individual's opinion on a controversial issue. Written language has also been used to determine an individual's level of expertise: Chujo and Utiyama (2005) constructed an extensive list of vocabulary along with the level of specialization that each word signaled. With this indicator, they found that an individual's written vocabulary could be evaluated to determine their level of specialization or education in a certain field. Klebanov and

Chujo & Utiyama's research provides strong evidence that written language can be used to predict future behavior. Subsequently, there is reason to believe that loan descriptions could reveal information about a borrower in the peer-to-peer lending setting.

There has been widespread interest in investing strategies on peer-to-peer lending sites and analyzing the tools investors have at their disposal, the most noteworthy being a borrower's credit score. In a 2009 analysis of 194,033 listings on Prosper Marketplace, researchers concluded that investors could only infer 33% of the difference in creditworthiness between two similar loans. They observed, The credit score provides an estimate of the true default probability, but it is only based on a subset of predictors. Despite this limitation, the credit score is the best available measure of the ex-ante default probability (Iver et. al., 2009). With the information that credit score was not the ultimate indicator of creditworthiness, peer-to-peer investors began searching for other ways to improve the overall return on their portfolio. The first analyst to consider descriptions as a potential indicator of creditworthiness was Peter Renton. Renton ran an analysis of description length versus default rate, and determined that loans with descriptions containing more than 2000 characters had a default rate approximately three times the average: 14.8% versus 4.6% (Renton, 2012). Although Renton has laid the groundwork for using descriptions as an indicator of likelihood for a borrower to default, his analysis was very shallow in that it did not take into account the content of the descriptions.

Previous research utilizing non-linear regression constructed a model that evaluated the correlation between word used by borrowers and default rates. When controlling for all other numerical variables, the words need, help, bills, and thank were found to be positively correlated with default, while credit, loan, and consolidate were negatively correlated with default (Wagers 2016). The most strongly correlated words indicated upwards of a .05% increased chance of default. While the appearance of a single word is indicative of a miniscule change in default risk, the collective effects of a borrower's entire vocabulary use could be a much stronger indicator.

Most, if not all research in the area of language, behavior, and peer-to-peer lending has taken a frequentist approach; however, the computer science field suggests a Bayesian approach may be more useful when analyzing natural language. A notable characteristic of Bayesian statistics, as opposed to a frequentist approach, is that it allows one to approach a question with an estimation, or prior belief about some data before any evidence is presented (Gelman 2002). In the last decade, the computer science branch of statistical language modeling (SLM) has been leaning towards Bayesian approaches. Rosenfeld (2000) gives a synopsis of what language analysis tools had been used in the previous decade, but then makes a strong case for Bayesian analysis as being the best way to approach natural language. Rosenfeld cites human biases as muddying the statistical process, saying: a better solution might be to encode such knowledge as a prior in a Bayesian updating scheme (Rosenfeld 2000). Rosenfeld sees the concept of the prior, specifically, to be the factor that gives Bayesian a leg up on other statistical approaches. More specifically, scholars have recognized the ability of the nave Bayes classifier as being simple yet accurate. Friedman (1997), in a comparison of numerous classification methods, concluded that, One of the most effective classifiers, in the sense that its predictive performance is competitive with state-of-the-art classifiers, is the so-called naive Bayesian classifier. After being established as an accurate classifier, McCallum and Nigam (1998) endorsed the use of the nave Bayes model specifically for text classification, citing that it was the most popular and consistently accurate classifier model. It is clear that the Bayesian approach works well with natural language, and will be the logical next step in providing further evidence linking natural language to creditworthiness in peer-to-peer lending.

IV. Hypothesis

The intention of this paper is to determine the extent to which a borrower's vocabulary can improve upon the estimation of their probability of default. My hypothesis is as follows:

Vocabulary will be a predictor of creditworthiness and will be able to improve upon the prediction accuracy of the credit score.

V. Data

The data itself consists of 36,055 loans from July 2007 to December 2011. Each loan is categorized as either Charged Off or Fully Paid. Notes that are still current have been excluded from the analysis. A loan is charged off when there is no longer a reasonable expectation of future payments and typically occurs when a loan is no later than 150 days past due (Lending Club, 2016). The data includes 66 pieces of information about each loan including information about the borrower (credit history, employment, income, location), the loan itself (amount, purpose, description submitted by borrower), and information that Lending Club has logged for the loan (interest rate, issue date, installment, loan status, past payments). The main interest here is of course the description submitted by a borrower describing why they are taking out this loan, and the specific words they use within that description that relate back to the hypotheses.

VI. Model

To test this hypothesis, I employ a nave Bayes classifier model. The nave Bayes gives us the probability of a classifier (i.e. the probability of default) given observed characteristics (predictors) about a certain instance (i.e. a given loan). The model estimates the probability of a loan defaulting with a multi-predictor rendition of Bayes' theorem [1], the formula for which is given by:

$$p(d|L) = p(c_1|d) * p(c_2|d) * \cdots * p(c_n|d)$$

[1]Bayes Theorem: $p(c_j|d) = p(d|c_j)p(c_j)p(d)$

Table 1: Descriptive Statistics

Variable	Mean	Median	Std. Dev.	Variance
Fully Paid	0.8499	-	0.3572	0.1276
Interest Rate	0.1177	0.1158	0.0365	0.0013
Prime Rate	0.0336	0.0325	0.0057	0.0000
"credit"	0.3624	-	0.4807	0.2311
"loan"	0.3860	-	0.4868	0.2370
"rate"	0.1604	-	0.3670	0.1347
"card"	0.3021	-	0.4592	0.2108
"interest"	0.1732	-	0.3785	0.1432
"payment"	0.2423	-	0.4285	0.1836
"finance"	0.0396	-	0.1949	0.0380
"consolidation"	0.0405	-	0.1971	0.0389
"consolidate"	0.1271	-	0.3331	0.1109
"financial"	0.0459	-	0.2093	0.0438
"need"	0.1169	-	0.3213	0.1033
"bills"	0.0955	-	0.2939	0.0864
"help"	0.1167	-	0.3210	0.1030
"you"	0.1798	-	0.3841	0.1475
"please"	0.0409	-	0.1980	0.0392
"thank"	0.1691	-	0.3748	0.1405
"monthly"	0.1158	-	0.3200	0.1024
"problem"	0.0212	-	0.1441	0.0208

Dummy variables for date, length of loan, income verification, location, length of employment, and loan purpose omitted from table. All words are dummy variables: 1 if yes, 0 if no

Where $p(d|L) =$ probability of loan L defaulting given its characteristics $c_1, c_2, \ldots, c_n$, and the probability of its characteristics being in a defaulted loan. For example, $p(c_1|d)$ gives the probability of characteristic one (c_1) being in a defaulted loan; perhaps the probability of the word need appearing in a defaulted loan as opposed to a non-defaulted loan.

To evaluate the additional predictive ability that vocabulary brings on top of traditional credit measures, I compare the prediction accuracy of two nave Bayes models: one that takes vocabulary into account and one that does not. The non-vocabulary model predicts loan default based on traditional credit measures, primarily contained in a given loan's interest rate. A loan's interest rate is primarily based on a borrower's credit score and thus reflects the probability of default for that borrower. Wagers (2016) showed that interest rate is primarily based on a borrower's FICO score, and also somewhat dependent on loan amount, income, and inquiries, which are not factors included in the FICO score but may affect risk on a loan. In terms of the composition of a FICO score, the algorithm is kept secret, but most believe that it is based upon the ratio of debt to available credit, which is in most cases a direct function of income. The score is then adjusted for payment history, number of recent credit applications, and negative events such as bankruptcy or foreclosure, as well as changes in income caused by changes in employment or family status (Arya et. al., 2011). This non-vocabulary model also uses a handful of other predictors including: date, location, length of loan, length of borrower's employment, prime rate at the time the loan was issued, whether the borrower's income was verified, and the purpose of the loan.

The model that contains vocabulary includes the same set of predictors as in the non-vocabulary model, complimented by a set of words of significance. Words of significance included in the model are split into two categories: positive words that are believed to decrease the probability of default and negative words that are expected to increase the probability of default. The sets mirror the words of significance used by Wagers (2016), and were chosen based on the hypothesis that they reflected either financial awareness (positive) or financial desperation (negative). The words were chosen in advance, before any observation of correlation with default rates. The sets are limited to avoid intercorrelation between words. The positive words in the model include: credit, loan, rate, interest, payment, finance, consolidation, consolidate, financial, while the negative words in the model include help, need, bills, you, thank, please, and problem. The model evaluates whether or not a borrower used each of these words in their loan description as an additional predictor towards whether a borrower will default or not. The construction of this second model allows for its direct comparison to the first model; any change in predictive ability can be attributed directly to the inclusion of vocabulary in the model.

Each of the models are trained on 80% of the 36,055 observations; on this portion of the data the models observe the probability of each characteristic being exhibited by a defaulted loan. They each take into account each of the variables mentioned above: the average interest rate for a defaulted loan, the location, etc, while the vocabulary model takes into account how likely each word is to appear in a loan that has been defaulted on. The models themselves provide conditional probabilities, essentially descriptive statistics, for each of its respective predictors including word frequency in fully paid and defaulted loans. The primary interest is using these models for prediction on the remaining 20% of the data. The models, having learned from training data, make predictions for whether each individual borrower defaulted or not on the test data. These predictions are then compared with the real-world outcomes for the same borrowers to determine the models' accuracy. However, splitting the dataset into 80% training data and 20% testing data means that the accuracy of the models for prediction will be somewhat dependent on how this split occurs. Thus, the partition into 80% training, 20% test data sets is repeated 10 times for each model, and results of the iterations averaged to provide a more accurate estimate of the predictive accuracy of each model. 10-fold cross validation is the minimum number of repetitions as recommended by Kohavi (1995) to provide a reasonable estimate of model accuracy.

It is necessary to address the fact that the nave Bayes model makes the nave assumption of conditional independence: that each feature $c_1, c_2, \ldots, c_n$ is independent of every other feature. It is this assumption that allows the model to achieve such simplicity, in that it allows each of the characteristics to be learned separately by the model. In this model, the implication is that all the characteristics of a loan are assumed to be independent of each other. Clearly, this is not the case in reality; loan characteristics are often related to other loan characteristics, and words in particular may

be conditionally dependent on each other due to context. Scholars have widely recognized that the nave Bayes model continues to perform very accurately even when breaking this assumption of conditional independence; consequentially it has been widely employed by academics in a variety of fields. McCallum and Nigam (1998) discuss this assumption: While this assumption is clearly false in most real-world tasks, naive Bayes often performs classification very well the function approximation can still be poor while classification accuracy remains high. In this paper I fully acknowledge the model breaking the conditional independence assumption, with the knowledge that it can still give an accurate prediction of loan default probability.

VII. Results

We can first use the models to evaluate conditional probabilities for the words of significance for defaulted and fully paid loans, respectively. These probabilities are shown in Tables 2 and 3.

The conditional probabilities generally suggest that words thought to signify financial awareness appear more frequently in fully paid loan descriptions, while words thought to signify financial desperation appear more frequently in loan descriptions in which the loan was eventually defaulted on. Particularly strong differences in frequency are shown in the word rate (found in 4% more fully paid loans than defaulted loans) and the word bills (found in 4% more defaulted loans than defaulted loans). These probabilities give a good insight into the links between vocabulary and default tendencies; however, comparing the predictive abilities of the Vocabulary-Included and Non-Vocabulary models will provide the true significance of these links. Below, tables are shown of the predictive results for 10 iterations of both the Vocabulary-Included and the Non-Vocabulary models.

Table 4 shows that the mean prediction accuracy of the Non-Vocabulary model after 10-fold cross validation is 79.7101%, with a similar median. A small standard error signifies the reliability of this mean. The individual iterations range from 78.9459% to 80.2913% accuracy; the details of each iteration can be found in Appendix Table A1. The sensitivity of the model is high, signifying that it is relatively good at determining when someone is going to pay his or her loan in full, while the specificity is low, signifying the model misses a high proportion of loans that were defaulted on in reality. Clearly there are some imperfections in the model, but this is not concerning given that I am conducting a model comparison, so both models will have the same imperfections.

Table 5 shows that the mean prediction accuracy of the Vocabulary-Included model after 10-fold cross validation is 80.1914%, with a similar median. Again, a small standard error signifies the reliability of this mean. The individual iterations range from 79.6949% to 80.5825% accuracy; the details of each iteration can be found in Appendix Table A2. The sensitivity and specificity of this model are relatively similar to those of the non-vocabulary model, reassuring us that the two models will share the same imperfections.

In a comparison of the two models, the Vocabulary-Included model predicts with a mean accuracy that is .4813% higher and a median accuracy that is .4715% higher than the accuracy of the Non-Vocabulary model. As displayed in Figure 1, the standard errors for the two models don't overlap, indicating that the differences in the means are statistically significant.

VIII. Discussion

The results have significant implications in terms of the original hypothesis that taking into account a borrower's vocabulary can improve the prediction of whether a borrower will default on a loan. The significance of the difference between the predictive abilities of the Vocabulary-Included model and the Non-Vocabulary model provides clear evidence for the hypothesis, suggesting that vocabulary did, in fact, improve the prediction of whether a borrower would default on a loan. The magnitude of the difference, approximately .48%, seems minuscule at first glance but is surprisingly large for a metric going unused in the financial industry. While this might not be the exact amount by which vocabulary improves the prediction of default, it signifies that there is potential for the use of vocabulary to improve the accuracy of currently used creditworthiness measures, such as the common credit score. Below are the confusion matrices of the two models and magnified to a sample size of 10,000 loans.

The confusion matrices imply that for a 10,000-loan sample, the Vocabulary-Included model will predict 9 more defaults than the Non-Vocabulary model, 819 to 810, and 39 more Fully Paid loans, 7161 to 7200, for a 48-loan improvement in loan accuracy overall. Although the same number of loans default in each case (1,834), an investor using the Vocabulary-Included model would be able to turn a more profitable investment from these loans, particularly because they would be able to foresee 9 more defaults that a traditional investor.

IX. Interpretation

The somewhat striking significance also brings up a question of whether vocabulary is acting as a proxy for another variable that is not included in the model: education. One could see this as a plausible explanation: with more years of education, a borrower is more likely to be considering the financial repercussions of their loan, and thus using more words like consolidate, credit, etc. Although no data exists on education levels for borrowers on Lending Club, many studies have proven education to be highly correlated with income (Porter, 2014). By controlling for income and every other available factor that lending club provides about borrowers, the model aims to eliminate the impact of a borrower's level of income to the best extent possible. Additionally, because the model evaluates only the frequency of certain words and not grammar, sentence structure, etc., education is less likely to be a confounding factor in the significance of vocabulary as a predictor of creditworthiness.

Table 2: Conditional probabilities for positive words of significance given default status

	credit	loan	rate	finance	consolidation	consolidate	financial
Charged Off (defaulted)	37.74%	37.24%	12.72%	2.90%	4.34%	11.73%	4.57%
Fully Paid	36.64%	38.79%	16.50%	4.05%	3.95%	12.69%	4.61%

Table 3: Conditional probabilities for negative words of significance

	need	bills	help	you	please	thank	monthly	problem
Charged Off (defaulted)	14.13%	12.65%	13.69%	18.45%	4.46%	17.87%	11.15%	2.17%
Fully Paid	11.42%	8.93%	11.25%	17.80%	3.98%	16.70%	11.63%	2.05%

Table 4: Results for the Non-Vocabulary model

Non-Vocabulary	Accuracy	95% CI Low	95% CI High	Sensitivity	Specificity	Precision
Mean	79.71%	78.76%	80.63%	87.48%	35.68%	88.51%
Median	79.74%	78.79%	80.66%	87.53%	35.63%	88.54%
Standard Error	0.12%	0.12%	0.12%	0.11%	0.30%	0.06%

Table 5: Results for the Vocabulary-Included model

Vocabulary-Included	Accuracy	95% CI Low	95% CI High	Sensitivity	Specificity	Precision
Mean	80.19%	79.25%	81.11%	87.86%	36.76%	88.72%
Median	80.21%	79.27%	81.12%	87.83%	36.78%	88.72%
Standard Error	0.07%	0.07%	0.07%	0.09%	0.41%	0.06%

Figure 1: Prediction accuracy comparison for Vocabulary-Included and Non-Vocabulary models.

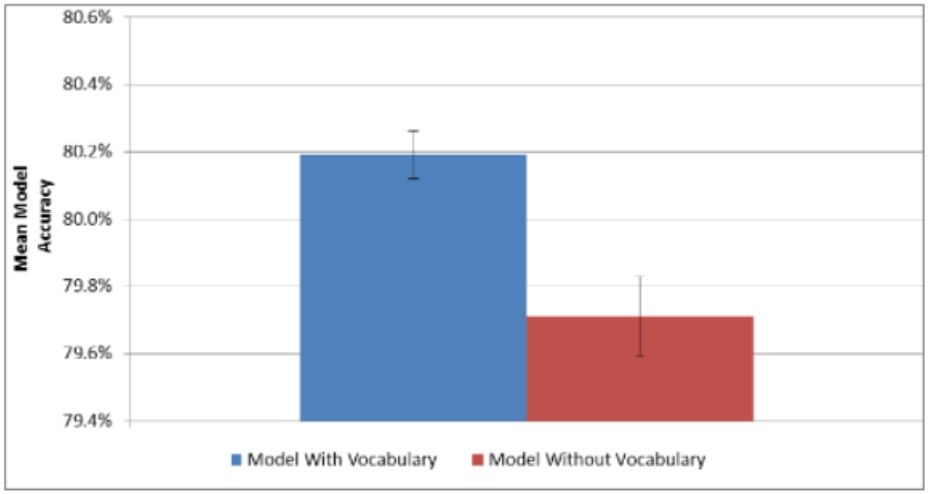

Table 6: Confusion matrix of Non-Vocabulary predictive model for theoretical sample size of 10,000 loans

Prediction	FALSE	TRUE
Default	1005	810
Fully Paid	1024	7161

Table 7: Confusion matrix of Vocabulary-Included predictive model for a theoretical sample size of 10,000 loans

Prediction	FALSE	TRUE
Default	966	819
Fully Paid	1015	7200

While there does seem to be a bright future for the use of vocabulary as a measure of creditworthiness, it also brings up an ethical question. If more weight is placed on vocabulary and similar metrics, it might have a disproportionate effect on low-income communities or certain racial groups. Under Title VII of the Civil Rights Act of 1964, employers are forbidden from using a racially neutral employment practice that has an unjustified adverse impact on members of a protected class. Thus, if vocabulary were adopted as a standard measure of creditworthiness in the financial industry, steps would have to be taken to ensure that its use did not have a disparate impact on any race, gender, etc. Kidder and Rosner (2002) showed how easy it was for even simple wording to have a disparate impact on certain ethnic groups, providing evidence that phrasing of SAT questions has a disproportionate impact on African-Americans. As more linguistic patterns are developed to predict creditworthiness, there will be more potential for some of these patterns to have a disparate impact on a protected class.

It is clear that we have not yet found the boundary of the predictive power of vocabulary. The significant results of this study provide momentum for future research on the predictive power of vocabulary, particularly in the financial industry. One direction this research could go would be to start analyzing more complexities of the vocabulary used by borrowers. This could include analysis of full phrases or misspellings as potential predictors of default. This research could also be extended to datasets from other peer-to-peer lending platforms and hopefully to a dataset from a larger financial institution, if privacy policies do not interfere.

These results may have implications beyond their use to analyze loan descriptions as well. If the usage of a few key words can be used as a signal of financial desperation, it is imaginable that similar patterns could be found within social media pages. In the future it may be possible to predict an individual's financial reliability by evaluating the vocabulary he or she uses on Facebook and elsewhere in the digital world. In fact, FICO, the leading company in credit scoring, is currently working on a way to assign a credit score to people who have been financially off the grid, meaning they have no credit history, by evaluating an individual's Facebook page (Selyukh, 2016). It has not been released whether FICO will be using users' vocabulary as a part of this evaluation, but from the results found in this paper, vocabulary patterns within these pages could be a useful tool. It seems that we are just beginning to scratch the surface when it comes to the predictive power of vocabulary.

X. Acknowledgements

Thanks to Garrett Milam for the continual support on this paper, as well as to Geoff Considine, America Chambers, and Lisa Johnson.

REFERENCES

[1] Arya, E., Arya, S., Wichman, C., & Wichman, C. (2013). Anatomy of the Credit Score. Journal of Economic Behavior & Organization,95, 175185.

[2] Athwal, N. (2014). The Disappearance Of Peer-To-Peer Lending. Retrieved from: http://www.forbes.com/sites/groupthink/2014/10/14/the-disappearance-of-peer-to-peer-lending/

[3] Bakker, E. (2016, April). Peer-to-peer lending markets: The leading countries for alternative finance and the next high-growth markets. Retrieved from http://www.businessinsider.com/peer-to-peer-lending-markets-the-leading-countries-for-alternative-finance-and-the-next-high-growth-markets-2016-4-26

[4] Boroditsky, L. (2009). How Does Our Language Shape the Way We Think? Retrieved from: https://www.edge.org/conversation/lera_boroditsky-how-does-our-language-shape-the-way-we-think

[5] Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. The American Economic Review, 103(2), 690-731.

[6] Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. System, 34(2), 255269.

[7] Civil Rights Act of 1964, Pub.L. 88-352, 78 Stat. 241 (1964).

[8] Freeman, J. (2012). The relationship between lower intelligence, crime and custodial outcomes: a brief literary review of a vulnerable group.Vulnerable Groups & Inclusion, 3(0).

[9] Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2-3), 131-163.

[10] Gelman, A. (2002). Prior distribution. Encyclopedia of environmetrics.

[11] Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2009). Screening in New Credit Markets: Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending? (SSRN Scholarly Paper No. ID 1570115). Rochester, NY: Social Science Research Network.

[12] Kidder, W. C., & Rosner, J. (2002). How the SAT Creates Built-in-Headwinds: An Educational and Legal Analysis of Disparate Impact. Santa Clara L. Rev., 43, 131.

[13] Klebanov, B., Beigman, E., and Diermeier, D. Vocabulary choice as an indicator of perspective. In ACL Short Papers, pages 253257, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[14] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2(12): 11371143

[15] Lending Club Investor Articles (2016). Retrieved from: http://kb.lendingclub.com/investor/articles/Investor/What-happens-when-a-loan-is-charged-off

[16] Lending Club Statistics. (2015). Retrieved from: https://www.lendingclub.com/info/download-data.action

[17] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).

[18] Mester, L. J. (1997). What's the point of credit scoring?. Business review, 3, 3-16.

[19] Pike, K. (1954). Language in relation to a unified theory of the structure of human behavior. The Hague, Netherlands: Mouton & Co.

[20] Porter, E. (2014). A Simple Equation: More Education = More Income. The New York Times. Retrieved from http://www.nytimes.com/2014/09/11/business/economy/a-simple-equation-more-education-more-income.html

[21] Renton, P. (2012). Loan Descriptions - Can They Be Helpful When Choosing Loans? Part 1. Retrieved from: http://www.lendacademy.com/lending-club-loan-descriptions-1/

[22] Rosenberg, J., & Tunney, R. J. (2008). Human Vocabulary Use as Display. Evolutionary Psychology, 6(3), 538549.

[23] Rosenfeld, R. (2000). Two Decades of Statistical Language Modeling: Where Do We Go From Here? Carnegie Mellon University Research Showcase.

[24] Trainor, S. (2015, July 16). The Long, Twisted History of Your Credit Score. Time. Retrieved from: http://time.com/3961676/history-credit-scores/

[25] Selyukh, A. (2016). Could Your Social Media Footprint Step On Your Credit History? Retrieved from: http://www.npr.org/sections/thetwo-way/2015/11/04/454237651/could-your-social-media-footprint-step-on-your-credit-history

[26] Wagers, J. (2016). Vocabulary as an indicator of creditworthiness: An analysis of public loan data. Working paper.

APPENDIX

Table A1: Cross-Validation for 10 iterations of the Non-Vocabulary model

Iteration	Accuracy	95% CI Low	95% CI High	Sensitivity	Specificity	Precision
1	79.8890%	78.9449%	80.8090%	87.5653%	36.1405%	88.6352%
2	79.6255%	78.6770%	80.5500%	87.2552%	36.4141%	88.5998%
3	79.5284%	78.5783%	80.4546%	87.3042%	35.4898%	88.4590%
4	80.0000%	79.0577%	80.9180%	87.7285%	36.2292%	88.6251%
5	80.2913%	79.3539%	81.2040%	88.2017%	35.4898%	88.5630%
6	79.7365%	78.7898%	80.6591%	87.5000%	35.7671%	88.5257%
7	80.0139%	79.0718%	80.9316%	87.5653%	37.2458%	88.7676%
8	78.9459%	77.9863%	79.8820%	86.9289%	33.7338%	88.1370%
9	79.7365%	78.7898%	80.6591%	87.5653%	35.3974%	88.4749%
10	79.3343%	78.3809%	80.2638%	87.1900%	34.8429%	88.3433%
Mean	79.7101%	78.7630%	80.6331%	87.4804%	35.6750%	88.5130%
Median	79.7365%	78.7898%	80.6591%	87.5326%	35.6285%	88.5443%
Standard Error	0.12%	0.12%	0.12%	0.11%	0.30%	0.06%

Table A2: Cross-Validation for 10 iterations of the Vocabulary-Included model

Iteration	Accuracy	95% CI Low	95% CI High	Sensitivity	Specificity	Precision
1	80.0277%	79.0859%	80.9452%	87.6469%	36.8762%	88.7182%
2	80.0971%	79.1565%	81.0133%	87.5653%	37.8004%	88.8558%
3	80.2080%	79.2693%	81.1223%	87.8916%	36.6913%	88.7169%
4	80.3745%	79.4386%	81.2857%	87.9243%	37.6155%	88.8669%
5	80.3051%	79.3681%	81.2176%	88.1201%	36.0444%	88.6408%
6	80.2497%	79.3116%	81.1632%	88.0059%	36.3216%	88.6715%
7	80.5825%	79.6502%	81.4900%	88.4628%	35.9519%	88.6654%
8	80.2080%	79.2693%	81.1223%	87.6795%	37.8928%	88.8834%
9	79.6949%	78.7475%	80.6182%	87.7774%	33.9187%	88.2672%
10	80.1664%	79.2270%	81.0814%	87.5326%	38.4473%	88.9552%
Mean	80.1914%	79.2524%	81.1059%	87.8606%	36.7560%	88.7241%
Median	80.2080%	79.2693%	81.1223%	87.8345%	36.7837%	88.7175%
Standard Error	0.07%	0.07%	0.07%	0.09%	0.41%	0.06%