



On the Improvement of Default Forecast Through Textual Analysis

Paola Cerchiello and Roberta Scaramozzino*

Department of Economics and Management, University of Pavia, Pavia, Italy

Textual analysis is a widely used methodology in several research areas. In this paper we apply textual analysis to augment the conventional set of account defaults drivers with new text based variables. Through the employment of *ad hoc* dictionaries and distance measures we are able to classify each account transaction into qualitative macro-categories. The aim is to classify bank account users into different client profiles and verify whether they can act as effective predictors of default through supervised classification models.

Keywords: text analysis, credit scoring, default, classification models, finance

1. INTRODUCTION

The change in all sectors of the economy that we are witnessing in recent years is so rapid that it speaks of the fourth industrial revolution. In the era of big data, across all sectors companies' main asset has become data. There is an increasing use of data, as the use of digital technologies increases, the amount of information collected increases exponentially. As a result, firms sit upon swathes of data, but the key is being able to derive value from them. In the financial sector, data are used for multiple purposes, one of which is credit scoring. This refers to the techniques used to assign creditworthiness to a customer to be able to distinguish "good" from "bad" customers i.e., clients who will repay their financing and those that will be insolvent. The probability that an applicant is insolvent is determined by analyzing the information on the latter (Hand and Henley, 1997). Credit scoring models are fundamental for banks to guarantee a correct forecast of default risk for financed loans, which translates into a reduction in losses and an increase in profits. There are numerous techniques for this purpose (Efron, 1979; Baesens et al., 2003; Jayasree and Balan, 2013; Emekter et al., 2015). Although nowadays most of the models in question use quantitative information, typically financial data, the latter is no longer sufficient to properly profile customers in a world that is now increasingly digital. This type of information, called hard information, is contrasted by another category of so-called soft information. Soft information is the term used to indicate information obtained through textual analysis, in this case, we talk about unstructured data. Text mining arises in this context. Text mining is the mechanism for extracting relevant information from unstructured documents to discover unknown patterns (Gupta and Lehal, 2009; Aggarwal and Zhai, 2012).

Even in today's standards, the traditional approach, which uses only hard information, is that which is widely used by firms but there is lack of studies that analyze textual information (Fei et al., 2015; Allahyari et al., 2017). Jiang et al. (2018) demonstrate how the use of textual data can increase the predictive power of a model, combining soft information with typically financial information analyzing the main p2p platforms in China. Groth and Muntermann (2011) state that exposure to intraday market risk management can be discovered through the use of text mining. Chan and Franklin (2011) show, through the use of textual data, that the forecast accuracy of their model improves similar traditional models by 7%. The advantages and disadvantages of Hard Information are analyzed by Liberti and Petersen (2018) who examine how information influences financial markets. Cornée (2019) demonstrates

OPEN ACCESS

Edited by:

Shatha Qamhieh Hashem,
An-Najah National University, Palestine

Reviewed by:

Yuze Yuan,
Minjiang University, China
Roberto Savona,
University of Brescia, Italy
Benjamin Miranda Tabak,
Fundação Getúlio Vargas, Brazil

*Correspondence:

Roberta Scaramozzino
roberta.scaramozzino01@
universitadipavia.it

Specialty section:

This article was submitted to
Artificial Intelligence in Finance,
a section of the journal
Frontiers in Artificial Intelligence

Received: 30 December 2019

Accepted: 13 March 2020

Published: 07 April 2020

Citation:

Cerchiello P and Scaramozzino R
(2020) On the Improvement of Default
Forecast Through Textual Analysis.
Front. Artif. Intell. 3:16.
doi: 10.3389/frai.2020.00016

the importance of textual information for credit scoring by analyzing 389 presences of a French bank. Grunert et al. (2005) analyzing German SMEs, compare a model based on financial data, one on textual data and one mixed demonstrates how the latter is the best in terms of predicting loan defaults.

In this paper, we propose a credit scoring model that utilizes text mining. The variables extracted through textual analysis are used as predictors in the model. To extract this information we have classified the bank transactions into macro-categories and then considered the frequencies of each macro category and the total amounts. We then compared the classical model based on financial information and the one with the addition of variables derived from textual analysis. The rest of the paper is organized as follows: in section Methodology the methodology used is shown, in section Data we analyze the data, in section Results we report the results obtained and finally the paper is discussed and future research presented.

2. METHODOLOGY

In this section we explain the methodology used. We developed a default risk prediction model by combining financial information and textual information. The first step of the analysis was the extraction of relevant variables from the texts provided in the transactions. The method consists of 2 parts: pre-processing and knowledge extraction. The first one needed a lot of work and demanded most of the time spent on the analysis. Preprocessing plays a significant role in Text mining. Any task of text mining fully depends on the preprocessing step. High-quality of preprocessing always yielded superior results (Kumar and Ravi, 2016). In the preprocessing, we have cleaned up the texts of the transactions. We have created the textual corpus starting from the original text through the removal of stop words (typical and specific for the context), tokenization, removal of errors, and stemming. We then created the document term-matrix. The matrix describes the frequency of the terms in the documents. In our case, the rows represent the transactions and the columns correspond to the terms. The columns are extracted through the analysis of the corpus regarding the dictionaries, these were created manually in terms of key-value: by inserting the subject of the service as a key and the category of the transaction as value. The final goal of text mining was to obtain topics (macro-areas of interest relating to transactions) to create new variables for each one created and to insert these variables in the credit scoring model. The first problem encountered is due to spelling errors, due to this problem many transactions were not found in dictionaries. This obstacle is because the transaction descriptions are handwritten by the bank operators, who often abbreviate the words or write absently without paying attention. To overcome such issue we, therefore, decided to use a distance measure, the Levenshtein one, to attribute the word to the closest transaction under consideration (Levenshtein, 1966). This is a measure accounting for the difference between two strings, which is the minimum number of changes necessary to transform one word into

another. Mathematically:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \left\{ \begin{array}{l} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{array} \right\} & \text{otherwise.} \end{cases} \quad (1)$$

The algorithm flow, accordingly, consists in:

1. Pre-processing on the single transactions and creation of a textual corpus;
2. Search in the dictionary of the key corresponding to the clean string obtained from preprocessing;
3. Assignment of the value corresponding to the key of the dictionary, that identifies the category. If this is not found, the Levenshtein distance of the string is calculated from all the keys in the dictionary and the nearest dictionary value is assigned, with a maximum threshold of 10. This means that whether the distance is greater, neither category will be assigned (at the end of the investigation, the percentage of Na is around 15%);
4. After assigning a category to each transaction of the dataset, we created the variables to be included in the credit scoring model.

The categories have been grouped into macro-categories, the macro-categories chosen for the analysis are 5:

- Non-essential goods: including expenses for goods such as shopping, travel and living.
- Essential goods: including expenses in markets, pharmacies.
- Financial services and utilities: including expenses related to banks, payment services, telco companies, petrol stations.
- Revenue: including incomes such as transfers and dividends.
- Salaries: including wages and pensions.

For each of the previous 5 variables, we have therefore created 2 further variables: frequency and the total amount spent by the client for the specific category. We have thus obtained 10 new variables through the processing of textual data.

The second step of the analysis is the application of credit scoring model. The textual based categories created were added to the financial ones and the new dataset was used in the model. The chosen model is the lasso logistics. Lasso logistic model is a shrinkage method that allows obtaining a subset of variables that are strongly associated with the dependent variable, through regularization of the coefficients bringing them to values very close or even exactly equal to zero. Since the L1 penalty is used, the variables with a coefficient equal to zero are excluded from the model (Hastie et al., 2009). Mathematically:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^n (y_i - x'_i \hat{\beta})^2 + \lambda \sum_{j=1}^m |\hat{\beta}_j|. \quad (2)$$

where y_i are the n-observations for the target variable (default/no default), x_i are the n-observations for the covariates, λ is the

penalization parameter chosen by cross validation and β are the coefficient of the model.

Along with Lasso we fitted Elastic net as well. Since there were no statistical significant differences, we preferred to focus on Lasso because of an easier interpretation of the results. Before applying the lasso logistic algorithm, we pre-selected the relevant variables through the Kruskal–Wallis test. We decided to apply this test due to the size of the dataset: too few observations (400) with respect to the number of available variables (52). When the size of available data is limited, as in our case, Lasso can be not efficient enough in fitting parameters. Lasso is good at dropping out not significant variables if it can use an appropriate number of observations compared to the number of variables. Thus, we pre-selected the most relevant variables (without being too restrictive) through Kruskal–Wallis paying attention to the division in training and test. Kruskal–Wallis is a non-parametric method (no assumptions on the distribution of the data) that states if there is a significant difference between the groups. The null hypothesis states that the k samples come from the same population and the alternative hypothesis states that they come from different ones (Siegel and Castellan, 1988).

The KW test (Conover, 1971) is the non-parametric version of the well-known ANOVA test and represents a multivariate generalization of the Wilcoxon test for two independent samples, that can be adapted to our problem as follows. On the basis of C independent samples (each containing the transactions of a client) of size n_1, n_C (the frequency of transactions for each client), a unique large sample L of size $N = \sum_{i=1}^C n_i$ is created by means of collapsing the original C samples. L can be organized as a matrix that contains a number of rows equal to N and a number of columns equal to W (the number of variables). Each entry of the matrix contains the frequency count of a specific variable along with each transaction. The KW test is then applied columns, in order to evaluate the discriminant power of each variable with respect to the client classification task. For each variable, the frequency vector corresponding to each column of the matrix L is ordered from the smallest frequency value to the largest one, and a rank is assigned to each transaction in the sample accordingly. Finally, one should calculate R_i as the mean of the ranks in each of the C original clients categories samples. The multivariate KW test can then be shown to take the following form:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^C n_i \frac{R_i^2}{n_i} - 3(n+1) \quad (3)$$

After having selected only the significant variables, we applied the lasso logistics comparing the model keeping only the financial variables, the model with only the textual variables and the one obtained by the combination of the variables. For each analysis, the dataset was divided into 2 parts, 70% for training and the remaining 30% for the test. In addition, the model has been cross-validated using 5 folds. We applied the cross validation in the training set and then validated in the test set.

The comparison of the 3 models is based on mean misclassification error (mmce), area under the curve roc (auc),

accuracy (acc), and roc curve (Krzanowski and Hand, 2009; Agresti and Kateri, 2011). Mmce is a prediction error metrics for a binary classification problem. The Roc curve is a graphical representation, along the two axes we find the sensitivity and 1-specificity, respectively represented by True Positive Rate and False Positive Rate. It is, therefore, the true positive rate as a function of the false positive rate. AUC is the area under the Roc curve, an aggregate measure of performance across all possible classification thresholds. Accuracy is the degree of correspondence of the estimated data with the real one.

3. DATA

We have undertaken this analysis starting from 2 datasets: loans and transactions relating to an Italian bank. The paper was executed in collaboration with Moneymour. Moneymour is a FinTech startup that offers a payment method to provide instant loans for online purchases. It allows client to buy immediately and pay in installments.

In the former the original variables were:

- Date of the loan request,
- Loan ID,
- Default status,
- Amount,
- Number and amount of loan payments.

In the latter, there were:

- Client,
- Accounting date,
- Value date,
- Transaction amount,
- Reason code and reason text.

This study analyzed 164931 transactions and 400 loans from 2015 to 2018.

The financial variables extracted are:

- Sum of income,
- Sum of outcome,
- Average income,
- Average outcome,
- Number of income,
- Number of outcomes,
- Total number of movements,
- Sum of salary and average salary.

All listed variables referring to the first month, three months, six months and the previous year respectively to the request for financing, and financing obtained.

Summary statistics of financial variables are reported in **Table 1**.

Through the use of text mining, the transactions carried out by each client were analyzed and the new variables were created:

- Salary,
- Total output non-essential goods,
- Total output essential goods,
- Total financial services and utilities,

TABLE 1 | Descriptive financial features.

	prev. funding	num. rev. month 1	num. rel. month 1	num. mov. month 1	num. rev. month 3	num. rel. month 3	num. mov. month 3	num. rev. month 6	num. rel. month 3	num. mov. month 6	num. rev. month 12	num. rel. month 12	num. mov. month 12
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.00	1.00	0.75	1.00	0.00	3.00	4.00	1.00	4.00	6.00	1.00	4.00	6.00
Median	0.00	1.00	6.00	7.00	3.00	16.50	21.00	6.00	34.00	39.00	11.00	55.00	71.00
Mean	0.23	1.46	13.78	15.23	4.22	41.38	45.59	8.10	81.39	89.50	14.93	150.50	165.40
3rd Qu.	0.00	2.00	23.25	26.00	6.00	74.25	83.25	12.00	149.25	165.20	22.00	248.20	271.80
Max.	0.00	9.00	90.00	93.00	43.00	253.00	265.00	59.00	466.00	485.00	97.00	873.00	911.00

- Total salaries,
- Total income,
- Frequency output non-essential goods,
- Frequency output essential goods,
- Frequency financial services and utilities,
- frequency salaries,
- Frequency income.

Summary statistics of textual variables are reported in **Table 2**.

The target variable is default or non-default of the client defined as follows: default means the non-fulfillment of loan payment installments for 3 months in a row.

4. RESULTS

In this section, we discuss the results obtained. The data set was divided into 2 portions: training and testing. Seventy percent of the data was used for training and the remaining 30% for the test. The data in both samples were distributed as follows: 37% defaulting and 63% non-defaulting. The target variable is the default status indicated with the value 0 for the non-default and with 1 the default.

We recall that the starting dataset presented 400 observations and 53 variables. The issue regarding the high number of variables with regards to the number of observations has been overcome by selecting the most significant variables through the Kruskal-Wallis test. The variables selected after applying the test are 21 and reported in the following list: previous funding, number revenue month 1, number releases month 1, number movements month 1, number revenue month 3, number releases month 3, number movements month 3, number revenue month 6, number releases month 6, number movements month 6, number revenue month 12, number releases month 12, number movements month 12, salary, total output essential goods, total financial services and utilities, frequency output non-essential goods, frequency output essential goods, frequency financial services and utilities, frequency salaries, frequency income. The model chosen for the credit scoring analysis is lasso logistics which represents an efficient choice in data analysis problems like ours when a variable selection step is needed.

For greater accuracy of the metrics obtained, we conducted the analysis using k-fold cross validation which partitions the dataset in subsets of equal size, where each subset is used as a test and the others as training. Moreover, we have conducted an out of sample analysis, training models on 2014-2015-2016 data and testing them on 2017 and 2018.

Parameters estimates of the 3 fitted logistic lasso models are reported in **Tables 3–5** referred respectively to financial variables, textual variables and the mixed one. In particular, from **Table 5** we can infer that several variables both financial and textual are significant. The textual ones, of course, are of major interest being the new ones. We observe that the largest parameter is obtained by the salary flag variable: having a negative sign means that the presence of the salary on the bank account decreases the probability of default. In particular if we calculate the odds ratio we get that the probability of non-defaulting is 12 times higher than the probability of defaulting. Thus, such

TABLE 2 | Descriptive textual features.

	Salary	Total output essential goods	Total financial services and utilities	Frequency output non essential goods	Frequency output essential goods	Frequency financial services and utilities	Frequency salaries	Frequency income
Min.	0.00	-31084	-224137.0	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.00	-1862	-19456.8	3.75	0.00	0.00	0.00	1.00
Median	0.00	0.00	-699.5	50.00	0.00	5.50	0.00	11.00
Mean	0.31	-2511	-15107.2	107.97	47.01	94.86	7.082	21.29
3rd Qu.	1.00	0.00	0.00	152.25	39.00	142.00	6.000	26.00
Max.	1.00	0.00	0.00	647.00	717.00	944.00	118.00	237.00

TABLE 3 | Important variables selected by Lasso Model on financial dataset.

Variables	Parameter estimated
Previous funding	-0.4548
Number revenue month 1	-0.1553
Number releases month 1	.
Number movements month 1	-0.1759
Number revenue month 3	.
Number releases month 3	-0.0000
Number movements month 3	-0.0150
Number revenue month 6	.
Number releases month 6	-0.0000
Number movements month 6	-0.0001
Number revenue month 12	0.0028
Number releases month 12	.
Number movements month 12	.

TABLE 4 | Important variables selected by Lasso Model on textual dataset.

Variables	Parameter estimated
Salary flag	-2.5998
Freq. salaries	.
Freq. income	-0.1111
Total output essential goods	0.0015
Total output financial services and utilities	0.0006
Freq. output non essential goods	-0.0083
Freq. output essential goods	.
Freq. output financial services and utilities	-0.1294

a simple information, that can be derived by the analysis of bank transactions, can add very useful information to the credit holder. Other two textual variables are worth of mentioning: frequency of income characterized by a negative sign and the total output for financial services and utilities with a positive sign. According to the former the larger is the frequency in the income the lower is the probability of default. On the other way around, a higher number of transactions for financial services and utilities increases the probability of default. This to say that having several incomes helps in affording financial loans but the impact of expenses for services and utilities is not negligible. Regarding purely financial variables, all of them but one shows negative signs, meaning that they reduce the probability of default. The three largest parameters are shown by “previous funding” “number of movements at month 1” and “number of revenue at month 1.” What affects largely the chance of repaying loans is the presence of previous loans request to the bank. This ensures a previous capability of respecting financial obligations. The same applies to the number of movements in the nearest month: having more movements can be considered a symptom of financial health as if we take into account the number of revenues.

Finally in **Table 6** we report the comparison among the 3 models: the first which considers only the financial variables, the second which considers only the textual variables and the third

TABLE 5 | Important variables selected by Lasso Model on mixed dataset.

Variables	Parameter estimated
Previous funding	-0.4984
Number revenue month 1	.
Number releases month 1	.
Number movements month 1	-.0.1650
Number revenue month 3	.
Number releases month 3	-0.0020
Number movements month 3	-0.0240
Number revenue month 6	.
Number releases month 6	.
Number movements month 6	.
Number revenue month 12	.
Number releases month 12	-0.0007
Number movements month 12	-0.0012
Salary flag	-2.5011
Freq. salaries	.
Freq. income	-0.0022
Total output essential goods	.
Total output financial services and utilities	0.0003
Freq. output non essential goods	.
Freq. output essential goods	.
Freq. output financial services and utilities	.

TABLE 6 | Results from Lasso logistic regression.

	Accuracy	Auc	False positive rate	Specificity
Financial features	0.75	0.931	0.479	0.521
Text features	0.84	0.939	0.233	0.767
Mix	0.80	0.946	0.356	0.644

one which combines both. The comparison is measured through Accuracy, Auc, type 1 error and specificity. As can be seen from **Table 6**, the results obtained are pretty high for the three models and even the roc curves tend to overlap many times as shown in **Figure 1**. If we focus on the accuracy, it is interesting to note that the model restricted to the textual variables component overcomes largely the model with only financial variables (84 vs. 75%). As a consequence the mixed model is an average of the two. On the other hand, the models are perfectly comparable in terms of AUC.

Nevertheless, there are important elements, first of all, classification errors. As you can see from **Table 6**, the mixed model and the text based one have an improvement over the type 1 errors. Not all errors have the same impact, some mistakes have higher implications than others. For a bank, type I error is the most dangerous one as it represents the probability of giving a loan to those who will not pay. The costs associated with type I errors are higher than type II errors.

5. CONCLUSIONS

In this paper, we use textual data to enhance the traditional credit scoring model. We evaluated the models performance by

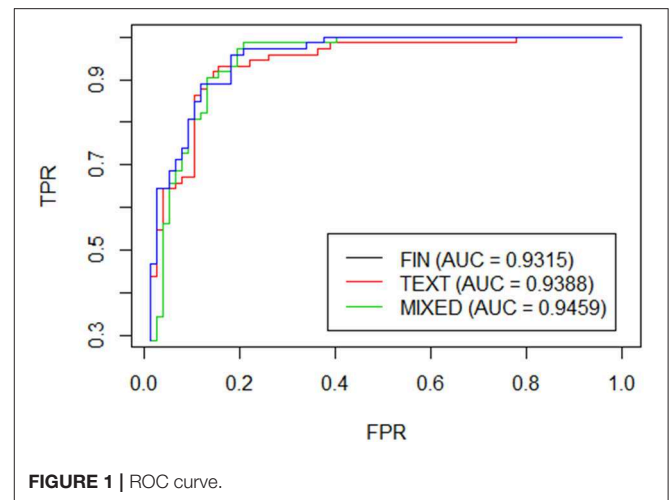


FIGURE 1 | ROC curve.

comparing the basic model in which only financial variables were included, against one in which there are only those extracted through text mining and the last one containing the mix of the two types of variables. From the analysis, we conclude that the addition of textual variables is relevant in the model. Although accuracy and AUC do not vary much, it should be emphasized first of all the distribution of errors. We observe an improvement over type 1 error and that both the textual based model and the mixed one show better results regarding accuracy. This is a promising result that encourages to further test the methodology with a larger dataset.

We can, therefore, state that despite the small size of the dataset, the analysis carried out shows how the textual analysis can be used in a credit scoring model to improve its accuracy.

Future research perspectives concern the application of the model to a larger dataset not only in terms of observations but also of variables based on textual information. More data can offer other types of information not available in the data at hand. Moreover the application of other text analysis technique like topic modeling (Cerchiello and Nicola, 2018) rather than the creation of manual dictionaries to make the process more automated and therefore decrease the time spent in pre-processing, can improve even more the quality of the analysis.

DATA AVAILABILITY STATEMENT

The data analyzed in this study was obtained from Moneymour and it is restricted by a non-disclosure agreement. Requests to access these datasets should be directed to RS, roberta.scaramozzino01@universitadipavia.it.

AUTHOR CONTRIBUTIONS

The paper is the product of full collaboration between the authors, however PC inspired the idea, the methodology and wrote sections Result and Conclusion. RS run the analysis and wrote sections Introduction, Methodology, and Data.

ACKNOWLEDGMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation program FIN-TECH:

REFERENCES

- Aggarwal, C. C., and Zhai, C. (2012). (ed.). *Mining Text Data*. Springer Science & Business Media.
- Agresti, A., and Kateri, M. (2011). *Categorical Data Analysis*. Berlin; Heidelberg: Springer.
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. et al. (2017). A brief survey of text mining: classification, clustering and extraction techniques. *arXiv:1707.02919*.
- Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *J. Operat. Res. Soc.* 54, 627–635. doi: 10.1057/palgrave.jors.2601545
- Cerchiello, P., Nicola, G. (2018). Assessing news contagion in finance. *Econometrics* 6:5. doi: 10.3390/econometrics6010005
- Chan, S. W. K., and Franklin, J. (2011). A text-based decision support system for financial sequence prediction. *Decis. Support Syst.* 52, 189–198. doi: 10.1016/j.dss.2011.07.003
- Conover, W. J. (1971). “The Kruskal-Wallis test,” in *Practical Nonparametric Statistics* (New York, NY: John Wiley and Sons Inc.), 256–264.
- Cornée, S. (2019). The relevance of soft information for predicting small business credit default: evidence from a social bank. *J. Small Bus. Manag.* (2019) 57, 699–719. doi: 10.1111/jsbm.12318e
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* 1, 1–26.
- Emekter, R., Tu, Y., Jirasakuldech, B., and Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Appl. Econ.* 47, 54–70. doi: 10.1080/00036846.2014.962222
- Fei, W., Gu, J., Yang, Y., and Zhou, Z. (2015). Credit risk evaluation based on social media. *Proc. Comp. Sci.* 55, 725–731.
- Groth, S. S., and Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decis. Support Syst.* 50, 680–691. doi: 10.1016/j.dss.2010.08.019
- Grunert, J., Norden, L., and Weber, M. (2005). The role of non-financial factors in internal credit ratings. *J. Bank. Fin.* 29, 509–531. doi: 10.1016/j.jbankfin.2004.05.017
- Gupta, V., and Lehal, G. (2009). A survey of text mining techniques and applications. *J. Emerg. Technol. Web Intell.* 1, 60–76.
- Hand, D. J., and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *J. R. Stat. Soc. Ser. A* 160, 523–541.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. London: Springer Science & Business Media.
- Jayasree, V., and Balan, R. V. S. (2013). A review on data mining in banking sector. *Am. J. Appl. Sci.* 10:1160. doi: 10.3844/ajassp.2013.1160.1165
- Jiang, C., Wang, Z., Wang, R., and Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Ann. Operat. Res.* 266, 511–529. doi: 10.1007/s10479-017-2668-z
- Krzanowski, W. J., and Hand, D. J. (2009). *ROC Curves for Continuous Data*. London: Chapman and Hall/CRC.
- Kumar, B. S., and Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowl. Based Syst.* 114, 128–147. doi: 10.1016/j.knosys.2016.10.003
- Levenshtein, V. I. (1966). “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet Physics Doklady*, 707–710.
- Liberti, J. M., and Petersen, M. A. (2018). Information: hard and soft. *Rev. Corporate Fin. Stud.* 8, 1–41. doi: 10.3386/w2507
- Siegel, S., and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences, 2nd Edn*. New York, NY: McGraw-Hill.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Cerchiello and Scaramozzino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.