Predicting Corporate Credit Ratings using Content Analysis of Annual Reports – A Naïve Bayesian Network Approach

Petr Hajek[1]✉, Vladimir Olej[1], Ondrej Prochazka[1],

[1] Institute of System Engineering and Informatics, Faculty of Economics and Administration, University of Pardubice, Studentska 84, 532 10 Pardubice, Czech Republic
{petr.hajek, vladimir.olej, ondrej.prochazka}@upce.cz

Abstract. Corporate credit ratings are based on a variety of information, including financial statements, annual reports, management interviews, etc. Financial indicators are critical to evaluate corporate creditworthiness. However, little is known about how qualitative information hidden in firm-related documents manifests in credit rating process. To address this issue, this study aims to develop a methodology for extracting topical content from firm-related documents using latent semantic analysis. This information is integrated with traditional financial indicators into a multi-class corporate credit rating prediction model. Informative indicators are obtained using a correlation-based filter in the process of feature selection. We demonstrate that Naïve Bayesian networks perform statistically equivalent to other machine learning methods in terms of classification performance. We further show that the "red flag" values obtained using Naïve Bayesian networks may indicate a low credit quality (non-investment rating classes) of firms. These findings can be particularly important for investors, banks and market regulators.

Keywords: Credit rating · Firms · Prediction · Concept extraction · Naïve Bayesian Network.

1 Introduction

Corporate credit ratings are intended to provide capital market participants with an evaluation for comparing the creditworthiness (capability and willingness of a firm to meet its payable commitments). The evaluation is particularly important for investors (institutional and individual), banks and market regulators, because it measures a default risk in a benchmark fashion. According to rating agencies such as Moody's, Standard & Poor's or Fitch, a credit rating is reported to require a variety of information necessary for the final evaluation. This information includes financial statements, corporate annual & quarterly reports, conference calls, management interviews, etc. The information is processed by a group of experts to reach an objective and independent rating grade (usually on a rating scale from Aaa/AAA denoting the highest credit quality to D representing default or bankruptcy).

In corporate default prediction literature, previous work has mainly focused on two approaches, structural and empirical [1]. The structural approach aims to model default probability based on the underlying dynamics of interest rates and firm-related indicators such as market capitalization [2]. In the empirical approach, on the other hand, the model is learned from data. The research has tended to focus either on the estimation of default probability or two-class bankruptcy prediction. This is mainly due to the specific characteristics of rating predictions such as the ordinal scaling of rating grades and multi-class prediction. Imbalanced classes are another issue to be addressed. As a result, it is difficult to measure the performance of prediction models.

The corporate credit rating is a time-consuming and expensive process, requiring an in-depth expert analysis of the underlying information. In recent years, there has been therefore an increasing interest in simulating the credit rating process of rating agencies through machine learning methods (e.g. [3-16]). These methods include hidden Markov models [3], neural networks [4,5], support vector machines [6-9], decision trees [10], fuzzy systems [11], rough sets [12], hybrid systems [13,14], and meta-learning approaches [15,16].

However, a major problem with this kind of application is the selection and accessibility of input variables (credit rating determinants). The main limitation of the above-mentioned studies is the focus on financial ratios (such as profitability, liability, or liquidity), which can be easily obtained from corporate financial statements.

Given the results of the studies (see e.g. [8] for a summary), it appears that the level of information available in financial data is bounded, resulting into a maximum of 80% accuracy for a multi-class problem [6]). This suggests that additional input variables are required to obtain significantly better results. This is also in line with the methodologies of rating agencies that emphasize the importance of qualitative factors in their credit rating process. Additionally, information extracted from firm-related textual documents have shown promising prediction ability recently. Specifically, the relative frequency of selected word categories in annual reports such as positive/negative sentiment [17] and modality/certainty/activity [18] have shown highly predictive abilities. Similarly, negative sentiment in news articles was reported more important for future credit rating changes compared with positive sentiment [19,20]. The research to date has tended to examine predefined word categories (dictionaries related to overall sentiment/opinion) rather than the topical content of textual documents. For related bankruptcy prediction problem, Cecchini et al. [21] extracted words with highest relative frequency from corporate annual reports and performed the detection of synonymous words using WordNet ontology. However, the above-mentioned studies failed to detect the structures and links between the concepts in firm-related textual documents. To bridge this gap, this study was aimed to develop a methodology for extracting topical content from firm-related documents. We developed this methodology to examine the importance of firm-related textual concepts in the highly imbalanced ordered multi-class problem of rating prediction.

This information is combined with traditional financial indicators to predict corporate credit rating. We demonstrate that although financial indicators are critical to predict rating grades, textual information may increase prediction performance. We believe that this approach may contribute to a greater understanding of the linguistic character of firm-related textual documents. In addition, the application of Naïve

Bayesian networks enables developing the "red flag" values of predictive variables indicating the presence of a low credit quality. In contrast to other machine learning methods, Naïve Bayesian networks can be considered as probabilistic white-box classifiers, facilitating the understanding of complex relationships within the data through probability distributions [22]. As far as we know, such probability distributions have not been reported in the literature. Subsequently, they can also be used to better model default probability in the structural models.

The remainder of this paper has been divided into four sections. The paper first gives an overview of the research methodology applied to predict corporate credit ratings. Specifically, section 2 lays out the theoretical foundations of textual content analysis and Naïve Bayesian networks, respectively. Section 3 describes the result of the content analysis of corporate annual reports. Section 4 provides the results of experiments and analyzes the performance of the proposed approach. Finally, section 5 concludes this paper and discusses its implications.

2 Research Methodology

The research methodology (depicted in Fig. 1) includes collection and pre-processing of both financial indicators and text information. The relevancy of pre-processed words in a particular document were obtained using a traditional *tf.idf* (term frequency weighted with inverse document frequency) weighting scheme, where the relative frequency of a word in a document is compared to the inverse proportion of the word over the entire corpus of documents [23]. The application of latent semantic analysis led to a lower-dimensional semantic space, where topic analysis of the corpus could be performed. Then, the two categories of variables (textual and financial) were integrated into one prediction model, which consisted of feature selection and classification into rating classes.

2.1 Financial Indicators

Rating agencies do not make the determinants of corporate credit rating public. However, their methodologies suggest that financial indicators represent important factors in the corporate credit rating process. In previous studies (see e.g. [10] for a review), broader categories such as profitability, liquidity, leverage, and market value ratios are usually considered as the most important financial ratios.

For this study, a set of 35 financial indicators was drawn from the Value Line Database and Standard & Poor's database for 557 U.S. firms (mining and financial companies were excluded from the dataset since they require specific financial indicators). As presented in Table 1, the set included: (1) size of firms; (2) corporate reputation; (3) industry membership; (4) profitability ratios; (5) activity ratios; (6) business situation; (7) asset structure; (8) liquidity ratios; (9) leverage ratios; and (10) market value ratios. Data for all the financial indicators were collected for the year 2010.

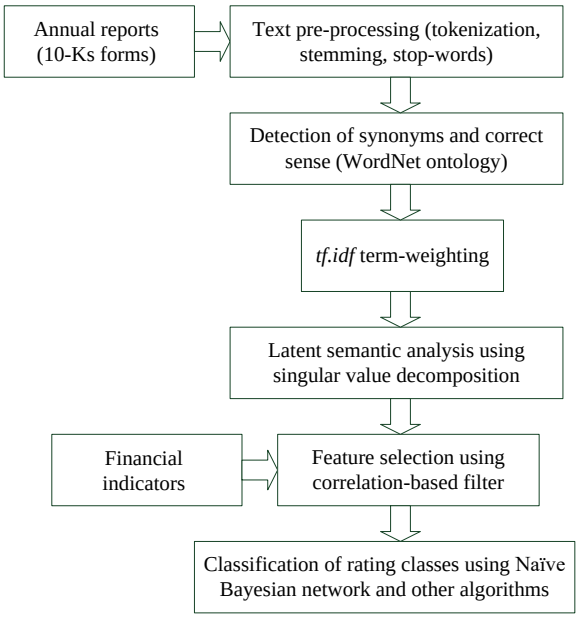

Fig. 1. Reseach methodology.

Table 1. List of financial indicators used in this study.

Category	Indicators
Size of firms	total assets, sales, cash flow, enterprise value
Corporate reputation	shares held by mutual funds, shares held by insiders
Industry membership	Standard Industrial Classification code
Profitability ratios	return on assets, return on equity, return on capital, operating margin, net margin, enterprise value/earnings before interest, taxes, depreciation and amortization
Activity ratios	sales/total assets, operating revenue/total assets
Business situation	effective tax rate, sales growth
Asset structure	share of fixed assets within total assets, share of intangible assets within total assets, non-cash working capital, working capital/total assets
Liquidity ratios	current ratio, cash ratio
Leverage ratios	book debt/total capital, market capitalization/total debts, market debt/total capital, net gearing
Market value ratios	dividend yield, 3-year stock price variation, beta, earnings per share, stock price/earnings, payout ratio, price-to-book value, high/low stock price

The firms were labelled with rating classes obtained from the Standard & Poor's rating agency in the year 2011. The rating classes are defined on the rating scale AAA, AA, … , D. Fig. 2 depicts the rating classes along with their frequencies in the dataset. Rating classes BBB, BB and B prevailed in the dataset, whereas rating classes C and D were not present at all. The frequencies also suggest a highly imbalanced classification problem.

Following recent studies on corporate credit rating prediction, we used feature selection procedure to include only informative financial indicators. Feature selection was also shown to improve the prediction performance of classification models in prior literature [10]. In order to provide the same subset of financial indicators for all classification algorithms, we used a correlation-based filter that optimizes the set of input variables by considering the individual predictive ability of each variables along with the degree of redundancy between the variables [24]. The correlation-based filter was chosen mainly because of the ordinal scaling of rating grades. Specifically, the rating grades were treated as the problem of ordinal classification. To avoid overfitting and feature selection bias, we used 10-fold cross-validation and performed the feature selection procedure only on training data, this is 10 times. All financial variables selected at least once are presented in Table 2 together with their mean values.

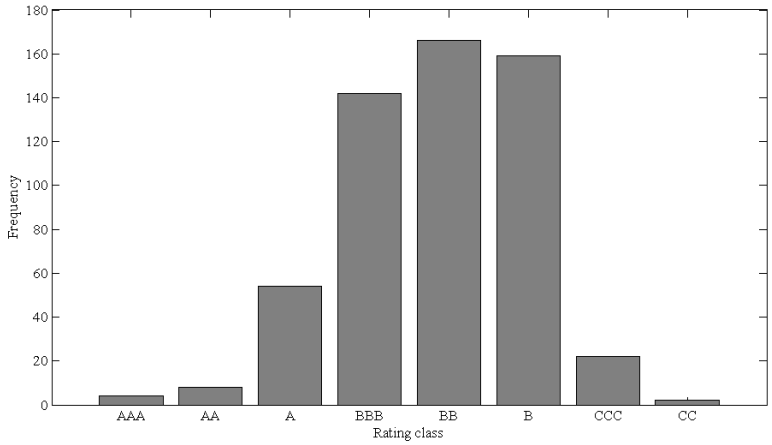

Fig. 2. Frequencies of rating classes in dataset.

Table 2. Mean values of selected financial indicators for rating classes.

Indicator	AAA	AA	A	BBB	BB	B	CCC	CC
Revenues	120.8	52.2	27.9	12.3	4.2	3.4	1.5	0.6
ROE	0.30	0.10	0.64	0.22	0.24	-0.03	-0.42	-0.51
MD/TC	0.03	0.14	0.16	0.26	0.34	0.53	0.60	1.00
EPS	3.64	4.73	3.25	3.09	2.04	0.85	-0.33	NA
High/Low	0.21	0.30	0.28	0.33	0.43	0.54	0.65	0.80
3yr stock var.	0.21	0.25	0.30	0.35	0.50	0.70	0.99	1.41
PR	0.41	0.53	0.52	0.99	0.40	0.31	0.04	NA
Dividend yield	0.03	0.04	0.03	0.06	0.02	0.02	0.02	0.00

Legend: ROE – return on equity, MD/TC – market debt to total capital, EPS – earnings per share, High/Low – high/low stock price, 3yr stock var. – 3-year stock price variation, PR – payout ratio, NA – missing value.

2.2 Latent Semantic Analysis for Concept Extraction

Documents are usually represented in a bag-of-words fashion (only the frequency of words matters, their order is ignored) with a very high dimensionality (each word representing one variable). However, a lower-dimensional semantic space is favorable for topic analysis. This dimensionality reduction can be performed using two general approaches, latent semantic analysis (using singular value decomposition - SVD) and probabilistic topic models (such as probabilistic latent semantic analysis or latent dirichlet allocation). We used latent semantic analysis in order to obtain an interpretable semantic model. In latent semantic analysis, semantic space is constructed from the SVD of the term-document matrix. In this new space, documents with the same concepts but different terms can be found [25].

SVD [26] is the factorization of the term-document matrix **X**, which have *m* lines (terms) and *n* columns (documents), into

$$\mathbf{X} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T, \tag{1}$$

where **U** (*m* x *m* dimension) and **V** is (*n* x *n* dimension) are orthonormal matrixes and **Σ** (*m* x *n* dimension) is diagonal (diagonal values are the singular values of the matrix **X**). The columns of **U** are the left singular vectors of the matrix **X**, and the columns of **V** (or the rows of $\mathbf{V}^T$) are the right singular vectors. To compute the SVD is to find the eigenvalues and the eigenvectors of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$, where the eigenvectors of $\mathbf{X}^T\mathbf{X}$ are the columns of **V** and the eigenvectors of $\mathbf{X}\mathbf{X}^T$ are the columns of **U**. The singular values of **X** in the diagonal of matrix **Σ** are the square root of the common positive eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$. The number of positive singular values equals the rank of the matrix.

In the term-document matrix **X**, it is important to select an appropriate term frequency weighting scheme because simply using the term frequency tends to exaggerate the contribution of the terms [25]. Commonly used term-weighting schemes, such as *tf.idf*, can address this issue.

2.3 Naïve Bayesian Networks

Naïve Bayesian Networks (also known as Bayesian Networks and Bayesian Belief Networks) are probabilistic graphical models that represent knowledge about an uncertain domain [27, 28]. Naïve Bayesian Networks consist of a set of nodes and set of directed edges between the nodes. Both the nodes and directed edges form a directed acyclic graph *G*. The nodes represent random variables. The edges represent direct dependences between the variables. The variables have a finite set of mutually exclusive states. All interdependencies are described using conditional probability distributions. To each variable with parents there is attached a probability table. Naïve Bayesian Networks are based on Bayes' theorem so that they can reason against the causal direction. Formally, a Naïve Bayesian Network *B* defines a unique joint probability distribution *P* over a set of random variables **U** [29] as follows:

$$P_B(X_1,...,X_n) = \prod_{i=1}^{n} P_B(X_i \mid \Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{X_i}}, \quad (2)$$

where $X_1, \ldots, X_n$ are random variables, and Θ represents the set of parameters that quantifies the network. Thus, independence assumption is encoded in graph G, this is each variable X_i is independent of its non-descendants given its parents in G.

Naïve Bayesian Networks are used to reason under uncertainty. They can estimate certainties for the values of variables that are not observed or their observation is very costly. They are also used as a representation for encoding uncertain expert knowledge in expert systems [30]. This is usually done by learning Naïve Bayesian Networks from data in order to induce a network that best fits the probability distribution over the set of training data. Heuristic search algorithms such as hill climbing, genetic algorithm or simulated annealing are used to find the optimum structure.

3 Content Analysis of Corporate Annual Reports

The annual reports (10-Ks forms) of corresponding 557 U.S. firms were collected at the U.S. Securities and Exchange Commission EDGAR System. The corpus of 557 filings was the result of document collection and pre-processing. The average document size (in number of characters) was 496,183. We downloaded all annual reports in txt format (without amended documents) for the year 2010. Documents that only referred to other reports were withdrawn. Similarly, graphics, tables and SEC header were removed from the documents before text pre-processing. First, all words were converted to lower case letters. Further, linguistic pre-processing included tokenization, stemming (Snowball stemmer) and discarding the stop-words in the corpus (using the Rainbow stop-word handler). Next, the potential term candidates were compared with the WordNet ontology [31] to detect synonyms and the correct sense of the terms for the domain (those with the highest score for Economy, Commerce or Law domains were chosen following [32]).

To represent the weights (term-weighting scheme) of the pre-processed words (i.e. how important a word is within a document), we used *tf.idf* as the most common approach. In this scheme, weights w_{ij} are calculated as follows:

$$w_{ij} = \begin{cases} (1+\log(tf_{ij}))\log\dfrac{N}{df_i} & \text{if } tf_{ij} \geq 1 \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

where N denotes the total number of documents, tf_{ij} is the frequency of the i-th word in the j-th document, and df_i denotes the number of documents with at least one occurrence of the i-th term. To select the most relevant words, we ranked them according to their *tf.idf* and used the top 1000 for our experiments. The most relevant 1000 words are usually enough to discriminate document categories from each other [33,34].

To extract the topical concepts from the corpus of corporate annual reports, we performed SVD and chose those concepts with singular values greater than 1 (76 concepts with the maximum singular value of 48.13), see Fig. 3. Further, the concepts had to be labeled based on the term importance. In the resulting vector space, semantic concepts can be interpreted due to the semantic relatedness between terms (they are placed near one another). Each term can be characterized by a weight indicating the strength of the semantic association. In other words, the concepts represent extracted common meaning components. Table 3 presents the concepts with the highest singular values along with the most important terms (largest weights). The meanings (labels) were manually assigned to the concepts based on the semantic association.

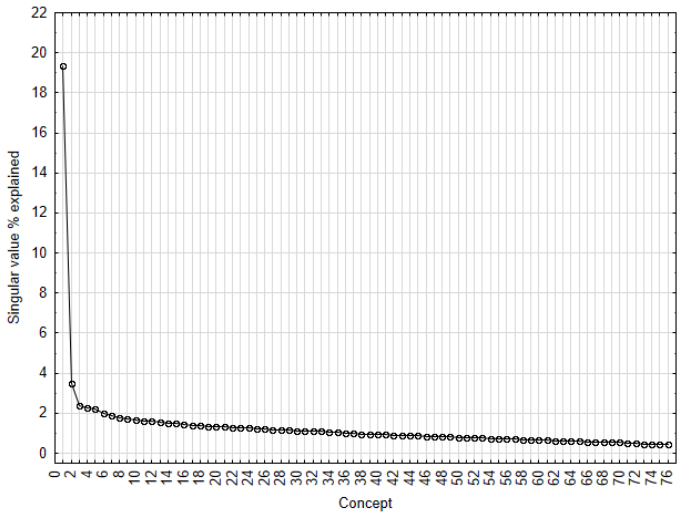

Fig. 3. Singular value explained by extracted topical concepts.

Table 3. Labels of topical concepts and representative words.

Label of concept	Most important words
Corporate restructuring	restructur, manufactur, swap, currenc, redempt, …
Relation to environment	manufactur, inventori, environment, labor, long-liv, …
Investment policy	indentur, indebted, construct, lender, libor, …
Financial restructuring	remedi, bond, court, alleg, lawsuit, …
Legal proceedings	court, alleg, licens, lawsuit, violat, …
Legal proc. implications	court, alleg, lawsuit, restructur, redempt, …
Debt policy	redempt, indentur, stock-bas, real, bond, …
Financial coop. and partnership	partner, currenc, merger, enterpris, third-parti, …
Foreign markets	polit, foreign, convert, countri, emerg, …
Domestic market difficulties	american, unfavor, cancel, forc, downturn, …
E-commerce	internet, space, billion, center, expans, …
…	…

Similarly as for the financial indicators, only informative concepts were used in subsequent corporate credit rating prediction. Therefore, the correlation-based filter was used to optimize the set of concepts. The following concepts were selected at least once: (1) corporate restructuring; (2) investment policy; (3) financial restructuring; and (4) domestic market difficulties. The mean values of the concepts for each rating class are presented in Table 4. These value suggest that firms with low credit quality mention corporate and financial restructuring less frequently in their annual reports. On the other hand, they used words related to investment policy and domestic market difficulties more frequently.

Table 4. Mean values of selected topical concepts for rating classes.

Concept	AAA	AA	A	BBB	BB	B	CCC	CC
Corporate restructuring	0.159	0.021	0.018	0.022	0.022	0.022	0.022	0.023
Investment policy	-0.024	-0.019	-0.014	-0.003	0.006	0.011	0.016	0.005
Financial restructuring	0.007	0.017	0.014	0.014	-0.004	-0.009	-0.011	-0.040
Domestic market difficulties	-0.021	-0.008	-0.011	-0.006	-0.001	0.004	0.008	-0.004

4 Experimental Results

To predict corporate credit rating using the combination of financial indicators and extracted concepts, we employed Naïve Bayesian network. To compare its performance, we performed the experiments also for several commonly used machine learning methods such as decision trees (C4.5 and Random Forest), neural networks (multilayer perceptron - MLP) and support vector machines (sequential minimum optimization algorithm - SMO), as well as statistical methods (logistic regression and k-nearest neighbor classifier). As stated above, we used 10-fold cross-validation to avoid over-fitting.

The methods were trained using the settings presented in Table 5. Naïve Bayesian network was trained using several heuristic search algorithms, namely a hill climbing algorithm, K2 (a hill climbing algorithm restricted by an order on the variables), genetic algorithm, and simulated annealing. Bayes scoring function was used to measure the quality of a network structure.

Common classification performance criteria such as accuracy may lead to misleading conclusions for imbalanced datasets [35]. Measures such as ROC (receiver operating characteristic) curve have been reported more appropriate for imbalanced datasets. We adopted this approach and measured the quality of prediction using the area under the ROC curve. A ROC is a graphical plot which illustrates the performance of a binary classifier system. Therefore, it is necessary to calculate average ROC across all classes to measure the overall performance of the methods. As reported by [36], the ROC measure has no obvious generalisation to multiple classes. However, it can be approximated by averaging the set of two-dimensional ROCs. Here we used a 1-vs-rest approach where ROC is weighted by class probability estimates [37].

Table 5. Settings of machine learning methods.

Method	Parameters and their values
Naïve Bayesian network	Hill climbing algorithm (no. of parents = {1, 2}) K2 (no. of parents = {1, 2}) Genetic algorithm (descendant population size = 10, population of network structures = 5, and no. of generations = 10) Simulated annealing (start temperature = 10, decreasing factor delta = 0.999, and no. of iterations = 10000)
C4.5	minimum no. of instances per leaf = 2, and confidence factor for pruning = 0.25
Random Forest	maximum depth of trees unlimited, no. of trees to be generated = {100, 200, 400}, and no. of variables randomly sampled as candidates at each split = $\log_2$(#predictors) + 1
SMO	complexity parameter C={2^0, 2^1, 2^2, … , 2^5}, polynomial kernel function with exponent = {1, 2}, RBF kernel function with gamma = 0.01
MLP	neurons in hidden layer = {2^2, 2^3, 2^4}, learning rate = 0.1, and no. of iterations = 500
Logistic regression	Broyden–Fletcher–Goldfarb–Shanno learning algorithm
k-nearest neighbor	k={3, 5} neighbors

Table 6 shows that Naïve Bayesian network performed statistically similar to Random Forest, SMO and MLP. The best network structure was found by the K2 hill climbing algorithm with 1 parent (represented by the output variable - rating class). In order to assess the impact of financial and textual indicators, additional experiments were conducted without considering these sets of variables. Table 7 presents the results obtained from the sensitivity analysis. Results from this table can be compared with the results in Table 6. As can be seen from Table 7, the performance of most classifiers significantly decreased when the financial indicators were not used. This result confirms their crucial importance in predicting corporate credit ratings. In contrast, no significant differences were found between the performances on all indicators and those on financial indicators (without textual indicators). However, the performance of the best classifiers (Naïve Bayesian network and Random Forest) decreased without using textual indicators, suggesting a limited information hidden in the text of corporate annual reports.

Table 6. Classification performance in terms of average ROC.

Method	Mean±St.Dev.	*t*-value (*p*-value)
Naïve Bayesian network	0.9237±0.0751	
C4.5	0.6051±0.2428	6.889 (0.000)*
Random Forest	0.9252±0.1747	-0.419 (0.676)
SMO	0.8702±0.2240	1.663 (0.100)
MLP	0.8754±0.2086	1.642 (0.105)
Logistic regression	0.8174±0.3005	2.280 (0.025)*
k-nearest neighbor	0.6204±0.2183	8.653 (0.000)*

Legend: * significantly worse than Naïve Bayesian network with all indicators at p=0.05 using Student's paired *t*-test.

Table 7. Classification performance in terms of ROC (Mean±St.Dev.) for datasets without financial and textual indicators.

Method	Without financial	Without textual
Naïve Bayesian network	0.7284±0.2208*	0.9221±0.0653
C4.5	0.5674±0.2458*	0.6365±0.2634*
Random Forest	0.6543±0.2832*	0.9188±0.1895
SMO	0.8621±0.0912	0.8691±0.2266
MLP	0.7608±0.3614*	0.9107±0.0912
Logistic regression	0.9066±0.2090	0.9207±0.0679
k-nearest neighbor	0.4929±0.0100*	0.7455±0.2528*

Legend: * significantly worse than Naïve Bayesian network with all indicators at p=0.05 using Student's paired t-test.

In Table 8, the probability distributions are presented for the input variables (note that only average values are shown across 10-fold cross-validation). The probabilities were merged to two values (> value / ≤ value) where more than two nodes were present for a variable. For example, ROE ≤ 0.128 is a "red flag" value, indicating the presence of a low credit quality (P=1-0.82=0.18 for AAA class, … , P=0.77 for CCC class and P=0.57 for CC class). Low rating classes (in this case BB, … , CC) are also known as non-investment rating classes. In fact, Table 8 indicates a strong change in probability distributions for this category of rating classes. Notably, the probability of non-investment rating class sharply increases for Revenues ≤ 8675, EPS ≤ 1.47, High/Low > 0.416, PR ≤ 0.168, dividend yield ≤ 0.008, financial restructuring ≤ 0.022 and domestic market difficulties > -0.010. Smaller changes can be observed inside the investment (AAA,…,BBB) and non-investment rating classes, respectively.

Table 8. Probability distribution for rating classes.

Variable	Value	AAA	AA	A	BBB	BB	B	CCC	CC
Revenues	>8675	0.82	0.79	0.62	0.34	0.08	0.05	0.02	0.14
ROE	>0.128	0.82	0.79	0.93	0.62	0.54	0.32	0.23	0.43
MD/TC	>0.312	0.18	0.21	0.11	0.36	0.47	0.83	0.77	0.57
EPS	>1.47	0.82	0.89	0.89	0.88	0.62	0.38	0.28	0.71
High/Low	>0.416	0.18	0.21	0.04	0.18	0.51	0.76	0.81	0.57
3yr stock var.	>0.327	0.25	0.35	0.40	0.56	0.86	1.00	0.98	0.88
PR	>0.168	0.90	0.72	0.75	0.73	0.31	0.13	0.07	0.17
Dividend yield	>0.008	0.90	0.72	0.88	0.80	0.43	0.32	0.33	0.50
Corporate restructuring	-	-	-	-	-	-	-	-	-
Investment policy	>-0.009	0.30	0.17	0.43	0.62	0.82	0.89	0.93	0.83
Financial restructuring	>0.022	0.30	0.39	0.30	0.37	0.10	0.03	0.07	0.17
Domestic market difficulties	>-0.010	0.10	0.39	0.59	0.58	0.72	0.86	0.76	0.83

Legend: - variable not selected as informative in Naïve Bayesian networks.

5 Conclusion

This paper has given an account of and the reasons for the widespread use of textual analysis in corporate credit rating prediction. Specific topics such as investment and

financial policy seem to be particularly important for credit rating assignment. The evidence from this study also suggests that capital market participants should pay attention to unusually low/high values of selected informative indicators.

The purpose of the current study was to design a methodology for extracting topical content from corporate annual reports. The methodology can also be applied to other firm-related documents such as earnings press releases, conference calls, news stories, analyst disclosures, and social media. Potential applications of topical content analysis include the prediction of future earnings, stock returns, volatility, financial fraud, etc. However, this study was limited to traditional latent semantic analysis mainly because we aimed to extract an easy-to-interpret semantic space. Probabilistic topic models, on the other hand, can be further extended to investigate other linguistic structures. Further investigation and experimentation into alternative topic models is therefore strongly recommended. For example, latent dirichlet allocation has recently been applied to extract topics from corporate press releases [38]. In addition, future studies should deal with the strong imbalance of credit rating datasets. Finally, to further our research we are planning to integrate the topic model with sentiment analysis to extract more informative indicators from firm-related documents.

The experiments in this study were carried out in Statistica 12 and Weka 3.7.13 using the MS Windows 7 operation system.

Acknowledgments. This work was supported by the scientific research project of the Czech Sciences Foundation Grant No: GA16-19590S and by the grant No. SGS_2016_023 of the Student Grant Competition.

References

1. Atiya AF (2001) Bankruptcy prediction for credit risk using neural networks: A survey and new results. IEEE Transactions on Neural Networks 12(4):929–935. doi: 10.1109/72.935101
2. Crouhy M, Galai D, Mark R (2000) A comparative analysis of current credit risk models. Journal of Banking & Finance 24(1–2):59–117, 2000. doi: 10.1016/S0378-4266(99)00053-9
3. Petropoulos A, Chatzis SP, Xanthopoulos S (2016) A novel corporate credit rating system based on Student's-t hidden Markov models. Expert Systems with Applications 53:87–105. doi: 10.1016/j.eswa.2016.01.015
4. Zhong H, Miao C, Shen Z, Feng Y (2014) Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. Neurocomputing 128:285–295. doi: 10.1016/j.neucom.2013.02.054
5. Hajek P (2011) Municipal credit rating modelling by neural networks. Decision Support Systems 51(1):108–118. doi: 10.1016/j.dss.2010.11.033
6. Huang Z, Chen H, Hsu CJ, Chen WH, Wu S (2004) Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision support systems 37(4):543–558. doi: 10.1016/S0167-9236(03)00086-1
7. Kim KJ, Ahn H (2012) A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. Computers & Operations Research 39(8):1800–1811. doi: 10.1016/j.cor.2011.06.023

8. Hajek P, Olej V (2011) Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning. Neural Computing and Applications 20(6):761–773. doi: 10.1007/s00521-010-0495-0
9. Chen CC, Li ST (2014) Credit rating with a monotonicity-constrained support vector machine model. Expert Systems with Applications 41(16):7235–7247. doi: 10.1016/j.eswa.2014.05.035
10. Hajek P, Michalak K (2013) Feature selection in corporate credit rating prediction. Knowledge-Based Systems 51:72–84. doi: 10.1016/j.knosys.2013.07. 008
11. Hajek P (2012) Credit rating analysis using adaptive fuzzy rule-based systems: an industry-specific approach. Central European Journal of Operations Research 20(3):421–434. doi: 10.1007/s10100-011-0229-0
12. Chen YS, Cheng CH (2013) Hybrid models based on rough set classifiers for setting credit rating decision rules in the global banking industry. Knowledge-Based Systems 39:224–239. doi: 10.1016/j.knosys.2012.11.004
13. Wu TC, Hsu MF (2012) Credit risk assessment and decision making by a fusion approach. Knowledge-Based Systems 35:102–110. doi: 10.1016/j.knosys.2012.04.025
14. Yeh CC, Lin F, Hsu CY (2012) A hybrid KMV model, random forests and rough set theory approach for credit rating. Knowledge-Based Systems 33:166–172. doi: 10.1016/j.knosys.2012.04.004
15. Pai PF, Tan YS, Hsu, MF (2015) Credit rating analysis by the decision-tree support vector machine with ensemble strategies. International Journal of Fuzzy Systems 17(4):521–530. doi: 10.1007/s40815-015-0063-y
16. Hajek P, Olej V (2014) Predicting firms' credit ratings using ensembles of artificial immune systems and machine learning - An over-sampling approach. In: Iliadis L, Maglogiannis I, Papadopoulos H (eds) IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 29–38, Springer, Berlin Heidelberg. doi: 10.1007/978-3-662-44654-6_3
17. Hajek P, Olej V (2013) Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. In: Iliadis L, Papadopoulos H, Jayne Ch (eds) International Conference on Engineering Applications of Neural Networks, pp. 1–10, Springer, Berlin Heidelberg. doi: 10.1007/978-3-642-41016-1_1
18. Hajek P, Olej V, Myskova R (2014) Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. Technological and Economic Development of Economy 20(4):721–738. doi: 10.3846/20294913.2014.979456
19. Lu YC, Shen CH, Wei YC (2013) Revisiting early warning signals of corporate credit default using linguistic analysis. Pacifin-Basin Finance Journal 24:1–21. doi: 10.1016/j.pacfin.2013.02.002
20. Lu HM, Tsai FT, Chen H, Hung MW, Li SH (2012) Credit rating change modeling using news and financial ratios. ACM Transactions on Management Information Systems 3(3):14. doi: 10.1145/2361256.2361259
21. Cecchini M, Aytug H, Koehler GJ, Pathak P (2010) Making words work: Using financial text as a predictor of financial events. Decision Support Systems 50(1):164–175. doi: 10.1016/j.dss.2010.07.012
22. Dejaeger K, Verbraken T, Baesens B (2013) Toward comprehensible software fault prediction models using Bayesian network classifiers. IEEE Transactions on Software Engineering 39(2):237-257. doi: 10.1109/TSE.2012.20
23. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5):513–523. doi: 10.1016/0306-4573(88)90021-0
24. Yu L, Liu H (2003) Feature selection for high-dimensional data: A fast correlation-based filter solution. In: International Conference on Machine Learning, ICML 2003, Washington, Vol. 3, pp. 856–863.

25. Crain SP, Zhou K, Yang SH, Zha H (2012) Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In: Aggarwal ChC, Zhai Ch (eds) Mining Text Data, pp. 129–161, Springer, New York. doi: 10.1007/978-1-4614-3223-4_5
26. Wall ME, Rechtsteiner A, Rocha LM (2003) Singular value decomposition and principal component analysis. In: Berrar DP, Dubitzky W, Granzow M (eds) A Practical Approach to Microarray Data Analysis, pp. 91–109, Kluwer. doi: 10.1007/0-306-47815-3_5
27. Howard RA, Matheson JE (2005) Influence diagrams. Decision Analysis 2(3):721–762. doi: 10.1287/deca.1050.0020
28. Pearl J (1988) Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann, San Mateo, CA.
29. Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 20(3):197–243. doi: 10.1007/BF00994016
30. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Machine Learning 29(2-3):131–163. doi: 10.1023/A:1007465528199
31. Miller GA (1995) WordNet: A lexical database for English. Communications of the ACM 38(11):39–41.
32. Hajek P, Olej V (2014) Comparing corporate financial performance and qualitative information from annual reports using self-organizing maps. In: 10th International Conference on Natural Computation (ICNC 2014), pp. 93–98, IEEE. doi: 10.1109/ICNC.2014.6975816
33. Matveeva I, Levow GA, Farahat A, Royer Ch (2007) Term representation with generalized latent semantic analysis. In: Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005, Current Issues in Linguistic Theory 292, pp. 45–54, John Benjamins Publishing.
34. Hajek P, Bohacova J (2016) Predicting abnormal bank stock returns using textual analysis of annual reports – A neural network approach. In: International Conference on Engineering Applications of Neural Networks (EANN 2016), pp. 67–78, Springer, Berlin Heidelberg. doi: 10.1007/978-3-319-44188-7_5
35. Chawla NV, Japkowicz N, Kotcz A (2004) Editorial: Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter 6(1):1–6.
36. Hand DJ, Till RJ (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45:171–186. doi: 10.1023/A:1010920819831
37. Provost F, Fawcett T (2001) Robust classification for imprecise environments. Machine learning 42(3):203–231. doi: 10.1023/A:1007601015854
38. Feuerriegel S, Ratku A, Neumann D (2016) Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In: Bui TX, Sprague RH (eds) 49th Hawaii International Conference on System Sciences (HICSS), pp. 1072–1081, IEEE. doi: 10.1109/HICSS.2016.137