News Sentiment in Bankruptcy Prediction Models: Evidence from Russian Retail Companies*

Elena Makeeva

Associate Professor

ORCID

E-mail: emakeeva@hse.ru

National Research University Higher School of Economics, Moscow, Russia

Maria Sinilshchikova

Researcher

ORCID

E-mail: mariasinilshchikova@gmail.com

Corporate Finance Center, National Research University Higher School of Economics, Moscow, Russia

News Sentiment in Bankruptcy Prediction Models: Evidence from Russian Retail Companies

Abstract

This study is aimed at investigating the application of news sentiment analysis to bankruptcy prediction models in the context of the Russian retail sector.

We analyse 190 companies: 95 Russian retail companies that went bankrupt in 2015-2019, and 95 non-defaulting analogue companies. This figure was attained from a larger pool of 312 companies retrieved from the Spark database on the basis of analysis of relevant financial data and further validated by the presence of pertinent news media coverage within 3 years of default date. The methodological base of this analysis is the logistic regression approach, used as a benchmark model, and several machine learning models: random forest, support vector machine, and multilayer perceptron.

The predictor set applied consists of 34 financial variables and sentiment variables, aggregated using the 'bag-of-words' from a total sample of 4877 news articles, from more than 800 distinct online resource locations. We establish a set of hypotheses based on a review of existing literature in the area, and evaluate their accuracy on the basis of our technical analysis.

Our results show that sentiment variables are statistically significant, and that adding sentiment variables improves the performance of bankruptcy prediction models. Also, the results indicate some reference characteristics of companies in terms of word-choice and descriptions in the news, indicating word choices correlated with financially stability and those correlated with financially instability.

Keywords: news sentiment, bankruptcy prediction models, random forest, support vector machine, multilayer perceptron

JEL classification: G17, G33

Introduction

Accurate analyses of a firm's financial stability are essential for multiple aspects of a company's planning and strategic processes, and are relied upon by other market participants, particularly banks. The probability of default, along with loss given default and exposure at default, are essential components in credit risk modelling. The problem of forecasting bankruptcy on the basis of financial reporting is connected with the fact that analysing the actual results of company reporting is possible only in the year subsequent to publication, which means that bankruptcy forecast in the short term is more challenging. Using news media resources for forecasting purposes assists with short term predictions by providing more current and recent data for analysis.

The COVID-19 pandemic of 2020 has caused the largest global recession in history. The overall effects of COVID-19 are not yet apparent, but it is already becoming clear the number of bankruptcies will rise enormously. The most novel and accurate methods of bankruptcy prediction are especially relevant today.

One of the main trends in bankruptcy prediction today is the application of new sources of information, will leads to an increase in the accuracy of the models. One of the most promising sources of information is textual data, which can be obtained from corporate disclosures, news, and social networks [1]. News sentiment analysis has been successfully used to predict stock price dynamics [2], but at the time of writing only a very small number of articles have been published dedicated to its application to the bankruptcy prediction problem. The advantage of using the news as a source of information is the availability and frequency of updated data, in comparison with traditional financial data and corporate disclosures sources, which are mostly updated once a year or quarterly [3].

Unlike the studies about the impact of the news sentiment on stock prices (where sentiment directly affects the value of shares of companies), the impact of news on the financial instability of a company should be understood as a description of the event which led to certain consequences for the company. To clarify, in the latter event, the news does indirectly affect the probability of default.

The methodological base of this work includes machine learning methods such as the random forest (RF) method, support vector machine (SVM) method, multilayer perceptron (MLP) method, and the logistic regression approach, which is to be used as a benchmark model. For the sentiment variable aggregation, 'the-bag-of-words' model [4] is used, along with the Linis Crowd dictionary [5].

Literature review

Machine learning in bankruptcy prediction models

Machine learning methods have received more attention than statistical methods, and in comparison with linear models, machine learning models provides higher accuracy.

As part of a review of 89 articles related to bankruptcy prediction models, Aziz [6] indicates an average accuracy ratio of 88% for machine learning and 84% for statistical models.

Also, machine learning methods do not have heavy restrictions on the entry data and thus are able to capture complex and non-linear patterns. On the other hand, the 'black box' results are not stable, have difficulties with interpretation, and trend towards overfitting [7]

According to Shi [8], the three most used methods of machine learning are:

1) Decision tree (DT). Unlike other machine learning methods, this is easy to interpret and can be displayed visually [9]. Tsai [10] claims that DT reaches the highest accuracy in comparison with other methods.

2) Artificial neural network (ANN), which shows a stable level of high performance in fitting nonlinear data, which in turn allows it to deal with very complex patterns. Ciampi [11] shows that ANN outperforms the Multiple Discriminant Analysis (MDA) and LR methods within a sample of 7000 companies, and is good for dealing with data omissions. However, ANNs tend to generalise results, which can lead to overfitting [12]. As such, Ding [13] claims that while ANN identifies only local optimum results, the support vector machine method (see below) succeeded in achieving global optimal results.

3) Support vector machine (SVM). Unlike ANN, SVM controls for errors with regard to generalising. SVM is successfully used in high dimensional nonlinear data and small datasets [14].

Sentiment analysis in finance

Sentiment analysis is a field of research based on methods of natural language processing, dedicated to identifying emotional attitudes either in relation to the subject under discussion in text to the object, or to a text as a whole.

One of the first studies about textual analysis in the field of finance belongs to Kohut [15]. The study suggests the content of the letters from companies' presidents letters differs between companies with high and poor financial performance.

According to Kearney [3], the sentiment analysis in finance can be divided into the following groups according to the source of information:

• Public corporate disclosures (annual reports and press releases) [16]

• The style and content of corporate disclosures signal as to the company's current situation and may contain useful information about future financial performance from the corporation's point of view. The limitation of this source of information is the low frequency of the data. The data is available only for the small number of the biggest companies and the disclosures are made on a quarterly or annual basis. Moreover, companies tend to try to manipulate public opinion to their benefit [17].

- Media, news articles and analysts' reports [2]
- Such compositions express observers' opinions about the overall financial and economic conditions, or about a particular industry or company. The advantage the news source is that news media and similarly published articles are available at all times and are frequently updated.
- Internet messages and social media networks [18]
- Social media networks are a potentially useful source of textual information because many people spend a considerable amount of time every day on the internet. However, internet messages, as opinions of common people, are among the noisiest sources of information because of the irrationality of such judgments and general unprofessionalism of internet users [3].
- Other or combined sources. [19]

The most common methods of sentiment analysis in finance are machine learning [16] and dictionary-based approaches [20].

According to the machine learning approach, the text is divided by tokens: e.g. by sentence, by word, or by combination of words. Each token is labelled with some category title. The machine learning algorithms predict these labels using the set of tokens.

The basis of the dictionary-based approach is a predefined dictionary with words arranged according to categories (e.g. positive and negative). Each word from the text is mapped with the dictionary word category. The dictionary-based method is associated with the 'bag of words' [4] because texts are considered to be unsorted sets of words.

The dictionary-based methods vary in how the dictionaries are defined and in how each word should be weighted.

The issue of the dictionary-based approach is that it is context-dependent: some words may have different tonalities in terms of different topics. This leads to the creation of topic-specific dictionaries, i.e. finance-specific dictionaries. Thus, the finance dictionary by [21] outperforms the traditional nonspecific Harvard dictionary in financial performance prediction and fraud detection [21].

Since the first wave of explosive interest in sentiment analysis, there has been a field of research on the sentiment of English language texts. However, no one has thus far succeeded in creating a successful multilanguage sentiment dictionary. The dictionary of Russian sentiment was created in 2016 by [5].

Each word found in such a dictionary may be weighted equally or have some weighting rule attached. The proportional scheme is called 'term frequency' (TF):

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{ij}}, \quad (1)$$

where – n_{ij} the frequency of the word i in the document j

k – number of documents

Another weighting is called 'inverse document frequency' (IDF):

$$IDF(w) = log\frac{N}{df_t} \quad (2)$$

where N– the number of all documents

df_t – the number of documents that contain the word w

Finally, term frequency – inverse document frequency (TF - IDF) is the result of multiplying (1) by (2):

$$TF\_IDF_{ij} = TF_{ij}IDF(w) \quad (3)$$

The main idea behind the IDF approach is that the most frequently occurring words are the least informative. [21] argues that the TF - IDF method outperforms the TF method. However, Azam [22] mentions that the TF method performs better on the smaller datasets. The article by Chen [23] suggests a more controlled term weighting method. Mai [24] uses TF - IDF weights for building a deep learning model.

Hypothesis

As a result of our literature review, the following hypotheses are thus articulated:

Hypothesis 1. *The TF - IDF word weights significantly increase bankruptcy prediction model performance in comparison with TF word weights.*

The TF - IDF is the more widely used weighting scheme [3], whereas TF weighting is more accurate in small datasets [22].

Hypothesis 2. *The number of news items has a statistically significant impact on the probability of default.*

To test how news items influence the probability of default, first we should check whether news coverage influences the probability of the default separately from the news content.

Hypothesis 3. *The news sentiment has a statistically positive impact on the probability of default. The application of the news sentiment variable significantly increases the model's performance.*

The evidence of the significance of textual sources other than news in financial insolvency predictions is provided in [25], and [24]

Hypothesis 4. *Negative news has a more significant effect on the probability of default than positive news does.*

The positive / negative influence of positive / negative news in the context of stock market activity is confirmed in articles [26], [27]. Leung [28] claims that positive news articles do have influence on the market. Apergis [29] shows that negative news articles influence the stock market more than positive ones.

Methodology and Data

For modeling bankruptcy in this study, we use four simple and effective methods that have already established in

the above-mentioned literature: logistic regression (LR); random forest (RF); support vector machine (SVM) and multilayer perceptron (MLP).

To evaluate the predictive performance of the machine learning analysis, the AUC-ROC performance measurement approach is preferred due to the balance between the true positive and the true negative rate [30].

To reduce the effect of the high variation between different splits and provide robust results, a 5-fold cross validation was conducted. Firstly, the sample was randomly split into 5 equal parts (subsample 1 (S1), … subsample 5 (S5). Then, S2 + S3 + S4 is used as the training set, while S1 is used as the test set. By repeating this step 5 times, each subsample is used as the test set one time. Thus, we get 5 AUC-ROC values and can calculate the mean AUC-ROC and standard deviation.

To test the significance of the model performance change, we conduct paired sample t-tests on the AUC-ROC metrics on the different splits and models, following the methodology of [24] and [31].

The analysis is performed in Python using the "scikit-learn" library.

Textual analysis methods

For our Hypothesis 1 (Model 1), the matrices of TF and TF-IDF frequencies of 658 dictionary words as columns and articles as rows are calculated (the 'bag-of-words' method). This array of columns represents the set of predictor variables. The matrix dimension is presented as (number of articles) x (number of dictionary words mentioned in articles). The default flags are duplicated for company, with the number of news articles greater than 1.

For Hypothesis 2 – 4 (Models 2-4) the news articles are aggregated by company and TF and TF-IDF statistics for all dictionaries are calculated (again, the «bag-of-words» method). The matrix dimension is represented as (number of companies) x (number of dictionary words mentioned in articles). Next, sentiment variables are calculated as the column sum of the TF and TF-IDF weight arrays. For each company, positive sentiment is the column sum of the weights (either TF or TF-IDF) of positive words; and negative sentiment is the sum of weights (TF or TF-IDF) of negative words multiplied by (-1); sentiment is the sum of positive and negative sentiment. For TF it simply takes the form:

$$sentiment_t f_i = \frac{number\ of\ positive\ words\ for\ company\ i - number\ of\ negative\ words\ for\ company\ i}{number\ of\ positive\ words\ for\ company\ i + number\ of\ negative\ words\ for\ company\ i} =$$

$$= pos_s entiment_i + neg_s entiment_i$$

(4)

For the lexical base for Russian sentiment analysis, the dictionary is obtained from the Linis Crowd dictionary [5], being the first Russian sentiment dictionary. Linis Crowd is a HSE open-source project that contains a sample of internet texts on socio-political topics with user ratings, and a sentiment dictionary based on these texts. The sentiment for each word is based on the average score, which is scaled from -2 to 2. After processing all ratings and deleting neutral words, 2719 words were left. The words with positive sentiment are considered as positive, the words with negative sentiment are considered negative. The lexical base was expanded with 186 antonyms, synonyms, and single-root words. The final dictionary consists of 2906 words in initial form (1027 positive and 1878 negative).

Database

The sample of companies was assembled from the Spark database, based on the following criteria. Russian companies from the retail sector with a default between 2015-2019 (and their non-default pairs) were selected. Financial data from one year before the default was considered. News data was taken for the three year period prior to the default date. The 'size of companies' metric was established in terms of 'micro', 'small', 'medium' and 'big' companies, with revenues of more than 50 million RUB. As a result, a sample of 312 (156 default and 156 non-default) companies was collected.

Textual factors

As a source of textual data, Yandex News was preferred to other news aggregators [32], e.g. Google News, Yahoo News, or databases, e.g. Spark or Thomson Reuters [33], for the following reasons:

1) It covers the entire observation period (2012-2019). For example, Google News and Spark store news only for the last year.

2) Yandex News aggregates a lot of online journals, even small regional ones, which results in a high level of news coverage for both small and big Russian private companies.

3) Yandex's advanced algorithms ensure high search relevance for Russian-language queries.

For each company in the sample the following web search query is completed:

1) Query: full company's name in Russian. If there are several companies with a similar name, a key word is added to the company's name (place of the registration or industry).

2) Options: The time period from 3 years before the date of default to the date of default.

The total number of articles is 4877, from more than 800 different online journals. News items were found for 95 company pairs out of 156 pairs.

Figure 1. The most frequent negative (left) and positive (right) words.

Text preprocessing is done using Python libraries: Natural Language Toolkit NLTK 3.4.5 [34], Pymorphy2 0.8 [35], and base Python libraries.

After preprocessing, all words are matched with the 'positive' and 'negative' dictionary categories.

The most frequent negative words and the most frequent positive words are presented in Figure 1. The word clouds are made with the Python library 'word cloud' resource. From Figure 1, we may see that the most frequent negative words are related to crime and legal issues.

Results

In general, the results are satisfactory. A high average accuracy of bankruptcy prediction is achieved, and some significant information is extracted from the news.

The implications for our hypotheses
are as follows:

Hypothesis 1. The data do not provide enough evidence towards Hypothesis 1

Hypothesis 2 The data do not provide enough evidence towards Hypothesis 1

Hypothesis 3 The hypothesis is not rejected. The news sentiment has a significantly negative impact on the probability of default and adding the news sentiment variable significantly increases the model performance.

Hypothesis 4. The hypothesis is not rejected. The negative news sentiment has greater impact on the probability of default than positive news does. Also, in case of SVM and MLP, adding the negative news sentiment variable results in a statistically higher average performance than adding a positive news sentiment variable does.

We use 34 financial control variables from Sample 2, and perform a one-factor analysis.

After all the transformations, the following factors are included in the model (Model 0):

Table 1. Model 0. The final set of control financial variables: Logistic regression

	Variable	Coefficient	p-value	Significance
X2	net income / revenue	-2.9022	0.0000	***
X25	cash and cash equivalents / current liabilities	-1.4209	0.0225	*
X30	payables / revenue	2.2413	0.0000	***
X33	revenue / mid-year inventories	2.728	0.0000	***

The variables that are included with negative coefficients reduce the risk of default (net income margin and cash ratio) and those that are included with a positive sign increase the risk of default (payables / revenue, and revenue / mid-year inventories).

For the benchmark performance, the machine learning analysis for Model 0 is performed:

The initial performance of the model is not very high (Table 2). This may be partly explained by the small sample size and the presence of both small and big companies. The articles with a comparably small dataset (e.g. of 240 variables and accounting-based variables [14]) and [36] with the sample of 107 default companies provide a comparable average AUC-ROC performance of 60-80%.

Table 2. Model 0. The final set of control financial variables. Machine learning analysis.

Model \ ROC-AUC	Mean	SD
Logit regression	0.75	0.06
Random forest	0.77	0.06
Support vector machine	0.76	0.06
Multilayer perceptron	0.76	0.03

Logistic regression analysis

The number of news items is not a significant variable (Table 3). So, **the data do not provide enough evidence in support of Hypothesis 2**. This model is excluded from further analysis in the next subchapter.

Table 3. Model 2. Logistic regression

	Variable	Coefficient	p-value	Significance
X2	net income / revenue	-2.9362	0.0000	***
X25	cash and cash equivalents / current liabilities	-1.5292	0.0154	*
X30	payables / revenue	2.1162	0.0000	***
X33	revenue / mid-year inventories	2.7295	0.0000	***
news_number	number of news article for each company	0.0054	0.1165	

Table 4. Model 3. Logistic regression

	Variable	Coefficient	p-value	Significance
X2	net income / revenue	-2.8297	0.0000	***
X25	cash and cash equivalents / current liabilities	-1.2658	0.0600	.
X30	payables / revenue	2.1130	0.0001	***
X33	revenue / mid-year inventories	2.5121	0.0001	***
sent*	Sentiment	-0.4778	0.0001	***

*sentiment variables are calculated for TF-IDF word weights due to the low statistical difference. For a more detailed explanation see the next subchapter.

Table 5. Model 4. The logistic regression

	Variable	coeff	p-value	coeff	p-value
X2	net income / revenue	-2.7874	0.0000***	-3.0572	0.0000***
X25	cash and cash equivalents / current liabilities	-1.1641	0.0728*	-1.6942	0.0109*
X30	payables / revenue	2.4754	0.0000***	1.5455	0.0064**
X33	revenue / mid-year inventories	2.8039	0.0000***	2.2797	0.0003***
pos_sent	Positive sentiment	-0.2789	0.1141		
neg_sent	negative sentiment			-0.7245	0.0003***

*sentiment variables are calculated for TF-IDF word weights due to the low statistical difference. For a more detailed explanation see the next subchapter.

News sentiment is a significant coefficient, so **the first part of Hypothesis 3 is not rejected** under the confidence level more than 99.99% (Table 4). The news sentiment has a significant negative impact on the probability of default, which corresponds to the supposition that the greater the number of positive words (as the characteristics of the positive events) and the less the number of negative words (as the characteristics of the negative events), the lower the probability of default.

Negative sentiment is a significant variable under the confidence level of 90%, whereas positive sentiment is not (Table 5). The absolute value of the coefficient in the logistic regression is more for the negative sentiment. The

results of Table 5 permit us to infer that our **Hypothesis 4 is not rejected.**

One of the main challenges of bankruptcy prediction is the increase in explanatory power of the models. In the next subchapter, the issue of model accuracy changes when news sentiment variables are investigated.

Machine learning analysis

The first step of the analysis is to compare the TF (Model 1a) and TF-IDF (Model 1b) performance. For this purpose, 658 unique dictionary words mentioned in the text are tested as the default predictors with TF and IDF weights.

Table 6. Model 1a and 1b. Machine learning analysis.

Model \ ROC-AUC	Model 1a		Model 1b		two-tailed t-test
	Mean	SD	Mean	SD	p-value
Logit regression	0.60	0.07	0.59	0.07	0.8764
Random forest	0.62	0.06	0.61	0.07	0.9353
Support vector machine	0.64	0.04	0.64	0.04	0.9518
Multilayer perceptron	0.62	0.07	0.64	0.04	0.6715

A paired sample t-test (Table 6) shows that the mean difference between the performance of Model 1b and 1a is not statistically different from zero, which **does not provide any evidence for the acceptance of Hypothesis 1**. As TF-IDF statistics are more widely used, in the subsequent models the aggregated sentiment variable with TF-IDF weighting scheme is used.

The average AUC-ROC performance (Table 6), is relatively low: the resulting performance of analysis of the total word collection is comparable to one-factor analysis for financial variables. However, combined with aggregated control variables, the text variable could improve the performance of the model and, more importantly, could lead to the following findings concerning word significance:

Table 7. List of 100 most significant negative words by content group

Group	Negative
Legal	банкротство, иск, уголовный, банкрот, суд, следствие, отсудить, расторгнуть, арбитражный, прокуратура, расследование, арбитраж, обанкротить, преследование, правоохранительный, банкротить, следователь, арест, банкротит, взыскание, осудить, уголовка, преступление, неправомерный, обанкротиться, обвинительный
Crime	мошенничество, хищение, нарушитель, арестовать, сокрытие, наказание, коррупция, незаконно, угонять, умышленный, взятка
Accident	взрыв, погибнуть, пострадавший, жертва, аварийный, гибель, захват, пострадать, разрыв
Debt	долг, взыскать, задолженность, убыточный, коллектор, неуплата, просрочка
Conflict	обмануть, подозревать, недобросовестный, заподозрить, дискриминация, конфликт, недостоверный, пожаловаться, ложный, эксплуатировать, требовать,
Prohibition	ограничить, запретить, ограничение, изъять
Low-quality	проблемный, неисправный, испортить, задерживать, заглохнуть, износ, опасный, невыполненный, некачественный, задержать
Other	снести, расторгнуть, ликвидация, опоздание, пропасть, уходить, негативный, закрывать, отказываться, горелый, недостаток, заставить, ликвидировать, потерять, уйти, самовольный, закрыться, остановка, реорганизация, специфический, выброс, дефицит

The presence of legal issues, crime, accidents, debt, conflicts, prohibitions, and low quality of delivered products or services in news articles correlates with companies having some serious problems that may lead to bankruptcy (Table 7):

Table 8. List of 100 most significant positive words by content group

Group	Positive
Innovation& Research& Strategy	прогресс,технический, стратегический, стратегия, презентовать, запускать,исследование, обновить, запустить, технология, развитие, инновация, обучать, информационный, интеллектуальный, развиваться, инициатива, внедрение, обучение, блокчейн, реформа, запуск, преобразовать, осваивать
Awards& Success	рекорд,награждать, номинация, победа, выигрывать, премия, награда, успех, успешный, процветать, заслуга,
Social activities	церемония, фестиваль, фотовыставка, форум, социальный, искусство, выставка, творчество, оздоровительный, празднование
Career events	трудоустройство,профориентация, карьерный
Partnership	сотрудничество, сотрудничать
High-quality	удобный, продуктивный, комфортный, удобно, шикарный, соответствовать, удовлетворить, популярный, оригинальный, уникальность, исключительный, актуальный, универсальный
Improvement	стабилизироваться, открыть, расширение, открыться, предотвращение, реконструкция, расширить, реконструировать, улучшить, защитить, оснастить, восстановление
Voluntary	льготный, добровольно, добровольный, бесплатный, дарить, благотворительный,
Other	семейный, солнечный, инвестиция, застраховать, семья, динамично, уверенность, динамичный, помощь, инвестировать, вечный, сладкий, выгодно, субсидия, выгода, инвестиционный, неплохо, шустрый

On the other hand, mentions of innovation, research activity, awards, social and career events, high-quality products or services, and declarations of new partnerships are correlated with a company in circumstances of financial stability (Table 8).

The second step is to check whether adding the news sentiment variable (Model 3) may significantly increase the model quality in comparison with the benchmark model with only control variables (Model 0):

Table 9. Model 3. Machine learning analysis

	Model 3		Model 0		one-tailed t-test
Model \ ROC-AUC	**Mean**	**SD**	**Mean**	**SD**	**p-value**
Logit regression	0.84	0.03	0.75	0.06	0.0167**
Random forest	0.86	0.03	0.77	0.06	0.0072***
Support vector machine	0.82	0.02	0.76	0.06	0.0335*
Multilayer perceptron	0.86	0.01	0.76	0.01	0.0001***

Adding sentiment variable to the control variables allows us to significantly increase the average AUC-ROC performance (Table 9), which indicates that **the second part of Hypothesis 3 is not rejected**. The RF reaches its highest accuracy (at 90%) in this particular area of analysis.

The last step of our analysis involves checking whether negative news sentiment and control variables (Model 4b) are correlated with a higher increase in the model quality than the positive news sentiment and control variables (Model 4a) do:

Table 10. Model 4a and 4b. The machine learning analysis

Model \ ROC-AUC	Model 4a		Model 4b		one-tailed t-test
	Mean	SD	Mean	SD	p-value
Logit regression	0.82	0.04	0.84	0.04	0.3638
Random forest	0.82	0.04	0.86	0.03	0.1804
Support vector machine	0.81	0.03	0.84	0.03	0.0894·
Multilayer perceptron	0.80	0.02	0.84	0.03	0.0600·

The results of one-tailed paired sample t-test (Table 6) are ambiguous: under the significant level of 10% the mean difference between Model 4b and 4a performance is statistically greater than zero only for SVM and MLP.

All methods of performance are nearly on the same level of accuracy, which is a surprising fact for the LR. The first explanation for this result is that machine learning algorithms have a lower performance at small datasets. Another explanation is that data-preprocessing and feature selection is done according to the logistic regression assumptions and specifics, which may increase the LR performance. The final possible explanation concerns the specifics of the machine model. As such, RF may perform worse than LR if the formula in model training contains a high proportion of essential predicting factors [37]. According to Salazar [38], RF and SVM perform at nearly the same level when the number of predictors is small. Gaudart [39] claims that the neural networks do not outperform linear regression in the case of normality, and the presence of homoscedasticity, the independence of the errors, and our preprocessing methodology, brings our data very close to the normal distribution. The overfitting effect of the small sample size was minimised by application of the 5-fold validation. The small number of predictors is regarded as harmonising in terms of establishing the comparability of the logistic regression and the machine learning method results.

References

1. Wang G, Chen G, Chu Y. A new random subspace method incorporating sentiment and textual information for financial distress prediction. Electronic Commerce Research and Applications. 2018 May 1;29:30-49.

2. Gupta K, Banerjee R. Does OPEC news sentiment influence stock returns of energy firms in the United States?. Energy Economics. 2019 Jan 1;77:34-45.

3. Kearney C, Liu S. Textual sentiment in finance: A survey of methods and models. International Review of Financial Analysis. 2014 May 1;33:171-85.

4. Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics. 2010 Dec 1;1(1-4):43-52.

5. Koltsova OY, Alexeeva S, Kolcov S. An opinion word lexicon and a training dataset for russian sentiment analysis of social media. Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE. 2016;2016:277-87.

6. Aziz MA, Dar HA. Predicting corporate bankruptcy: where we stand?. Corporate Governance: The international journal of business in society. 2006.

7. Alaka HA, Oyedele LO, Owolabi HA, Kumar V, Ajayi SO, Akinade OO, Bilal M. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert Systems with Applications. 2018 Mar 15;94:164-84.

8. Shi Y, Li X. An overview of bankruptcy prediction models for corporate firms: A Systematic literature review. Intangible Capital. 2019;15(2):114-27.

9. Kumar PR, Ravi V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. European journal of operational research. 2007 Jul 1;180(1):1-28.

10. Tsai CF, Hsu YF, Yen DC. A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing. 2014 Nov 1;24:977-84.

11. Ciampi F, Gordini N. Managerial characteristics and default prediction modeling for small firms: Evidence from small Italian firms. In35 Convegno Annuale AIDEA-Management Senza Confini. Gli Studi di Management Tradizione e Paradigmi Emergenti 2012 (pp. 1-19). IT.

12. Chen N, Ribeiro B, Chen A. Financial credit risk assessment: a recent review. Artificial Intelligence Review. 2016 Jan 1;45(1):1-23.

13. Ding Y, Song X, Zen Y. Forecasting financial condition of Chinese listed companies based on support vector machine. Expert Systems with Applications. 2008 May 1;34(4):3081-9.

14. Chandra DK, Ravi V, Bose I. Failure prediction of dotcom companies using hybrid intelligent techniques. Expert Systems with applications. 2009 Apr 1;36(3):4830-7.

15. Kohut GF, Segars AH. The president's letter to stockholders: An examination of corporate

communication strategy. The Journal of Business Communication (1973). 1992 Jan;29(1):7-21.

16. Park S, Lee J, Kim K. Semi-supervised distributed representations of documents for sentiment analysis. Neural Networks. 2019 Nov 1;119:139-50.

17. Callison C, Zillmann D. Company affiliation and communicative ability: How perceived organizational ties influence source persuasiveness in a company-negative news environment. Journal of Public Relations Research. 2002 Mar 1;14(2):85-102.

18. Broadstock DC, Zhang D. Social-media and intraday stock returns: The pricing power of sentiment. Finance Research Letters. 2019 Sep 1;30:116-23.

19. Kothari SP, Li X, Short JE. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. The Accounting Review. 2009 Sep;84(5):1639-70.

20. Loughran T, McDonald B. Measuring readability in financial disclosures. The Journal of Finance. 2014 Aug;69(4):1643-71.

21. Loughran T, McDonald B. When is a liability not a liability. Journal of Finance, forthcoming. 2009 May 6.

22. Azam N, Yao J. Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Systems with Applications. 2012 Apr 1;39(5):4760-8.

23. Chen et al., "Turning from TF-IDF to TF-IGM for term weighting in text classification," Expert Systems with Applications, vol. 66, pp. 245-260, 2016

24. Mai F, Tian S, Lee C, Ma L. Deep learning models for bankruptcy prediction using textual disclosures. European journal of operational research. 2019 Apr 16;274(2):743-58.

25. Hájek P, Olej V, Myskova R. Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. Technological and Economic Development of Economy. 2014 Oct 2;20(4):721-38.

26. Caporale GM, Spagnolo F, Spagnolo N. Macro news and stock returns in the Euro area: a VAR-GARCH-in-mean analysis. International Review of Financial Analysis. 2016 May 1;45:180-8.

27. Dzielinski M. News sensitivity and the cross-section of stock returns. NCCR FINRISK. 2011(719).

28. Leung DK, Lee FL. How journalists value positive news: The influence of professional beliefs, market considerations, and political attitudes. Journalism Studies. 2015 Mar 4;16(2):289-304.

29. Apergis N, Miller SM. Consumption asymmetry and the stock market: empirical evidence. Economics Letters. 2006 Dec 1;93(3):337-42.

30. Olson DL, Delen D, Meng Y. Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems. 2012 Jan 1;52(2):464-73.

31. Gardner J, Brooks C. Statistical Approaches to the Model Comparison Task in Learning Analytics.

32. Lu YC, Shen CH, Wei YC. Revisiting early warning signals of corporate credit default using linguistic analysis. Pacific-Basin Finance Journal. 2013 Sep 1;24:1-21.

33. Papakyriakou P, Sakkas A, Taoushianis Z. Financial firm bankruptcies, international stock markets, and investor sentiment. International Journal of Finance & Economics. 2019 Jan;24(1):461-73.

34. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."; 2009 Jun 12.

35. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. InInternational Conference on Analysis of Images, Social Networks and Texts 2015 Apr 9 (pp. 320-332). Springer, Cham.

36. Geng R, Bose I, Chen X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining. European Journal of Operational Research. 2015 Feb 16;241(1):236-47.

37. Kirasich K, Smith T, Sadler B. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. SMU Data Science Review. 2018;1(3):9.

38. Salazar DA, Vélez JI, Salazar JC. Comparison between SVM and logistic regression: Which one is better to discriminate?. Revista Colombiana de Estadística. 2012 Jun;35(SPE2):223-37.

39. Gaudart J, Giusiano B, Huiart L. Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data. Computational statistics & data analysis. 2004 Jan 28;44(4):547-70.

40. Engelberg J. Costly information processing: Evidence from earnings announcements. InAFA 2009 San Francisco meetings paper 2008.

41. Ahmadi Z, Martens P, Koch C, Gottron T, Kramer S. Towards bankruptcy prediction: deep sentiment mining to detect financial distress from business management reports. In2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) 2018 Oct 1 (pp. 293-302). IEEE.

42. Beaver WH. Financial ratios as predictors of failure. Journal of accounting research. 1966 Jan 1:71-111.

43. Fulmer JG, Moon J, Gavin TA, Erwin M. A bankruptcy classification model for small firms. Journal of commercial bank lending. 1984 Jul;66(11):25-37.

44. Fedorova E, Demin I, Khrustova L, Osetrov R, Fedorov F. The influence of the tone of the CEO's letters on the company's financial performance. Russian Journal of Management. 2017;15(4).

45. Ohlson JA. Financial ratios and the probabilistic prediction of bankruptcy. Journal of accounting research. 1980 Apr 1:109-31.

46. Qu Y, Quan P, Lei M, Shi Y. Review of bankruptcy prediction using machine learning and deep learning techniques. Procedia Computer Science. 2019 Jan 1;162:895-9.

47. Bellovary JL, Giacomino DE, Akers MD. A review of bankruptcy prediction studies: 1930 to present. Journal of Financial education. 2007 Dec 1:1-42.

48. West RC. A factor-analytic approach to bank condition. Journal of Banking & Finance. 1985 Jun 1;9(2):253-66.

49. Merton RC. On the pricing of corporate debt: The risk structure of interest rates. The Journal of finance. 1974 May 1;29(2):449-70.

50. Feller W. An introduction to probability theory and its applications, vol 2. John Wiley & Sons; 2008.

51. Nilsson P. The influence of related and unrelated industry diversity on retail firm failure. Journal of Retailing and Consumer Services. 2016 Jan 1;28:219-27.

52. Gleason KC, Mathur LK, Mathur I. The interrelationship between culture, capital structure, and performance: evidence from European retailers. Journal of business research. 2000 Nov 1;50(2):185-91.

53. Ahmad Z, Abdullah NM, Roslan S. Capital structure effect on firms performance: Focusing on consumers and industrials sectors on Malaysian firms. International review of business research papers. 2012 Jul;8(5):137-55.

54. Gaur V, Fisher ML, Raman A. An econometric analysis of inventory turnover performance in retail services. Management science. 2005 Feb;51(2):181-94.

55. Akinboade OA, Mokwena MP, Kinfack EC. Estimating profit efficiency among South African retail firms using stochastic frontier approach. The International Review of Retail, Distribution and Consumer Research. 2008 Dec 1;18(5):511-20.

56. Makeeva E, Bakurova A. Forecasting bankruptcy oil and gas companies using neural networks. Journal of Corporate Finance Research. 2012(3 (23)):22-30.

57. Демешев ББ, Тихонова АС. Динамика прогнозной силы моделей банкротства для средних и малых российских компаний оптовой и розничной торговли. Корпоративные финансы. 2014;8(3).

58. Mselmi N, Lahiani A, Hamza T. Financial distress prediction: The case of French small and medium-sized firms. International Review of Financial Analysis. 2017 Mar 1;50:67-80.

59. Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The journal of finance. 1968 Sep 1;23(4):589-609.

60. Altman EI, Sabato G. Modeling credit risk for SMEs: Evidence from the US market. InManaging and Measuring Risk: Emerging Global Standards and Regulations After the Financial Crisis 2013 (pp. 251-279).

61. Pradhan R, Pathak KK, Singh VP. A Study of Credit Lending and Prediction of Bankruptcy using Back-Propagation Neural Networks. IIMS Journal of Management Science. 2012;3(2):179-93.

62. Zhang W. R&D investment and distress risk. Journal of Empirical Finance. 2015 Jun 1;32:94-114.