Bankruptcy Prediction Using Machine Learning Models with the Text-based Communicative Value of Annual Reports

Tsung-Kang Chen*
Email: vocterchen@nycu.edu.tw
Department of Management Science,
National Yang Ming Chiao Tung University, TAIWAN

Hsien-Hsing Liao
E-mail:hliao@ntu.edu.tw
Department of Finance,
National Taiwan University, TAIWAN

Geng-Dao Chen
Email: gengdaochen@gmail.com
Department of Management Science,
National Yang Ming Chiao Tung University, TAIWAN

Wei-Han Kang
Email: hugo.kang569@gmail.com
Department of Management Science,
National Yang Ming Chiao Tung University, TAIWAN

Yu-Chun Lin
Email: superrock.lin@gmail.com
Department of Management Science,
National Yang Ming Chiao Tung University, TAIWAN

* Corresponding author.

Bankruptcy Prediction Using Machine Learning Models with the Text-based Communicative Value of Annual Reports

Abstract

We investigate whether including the text-based communicative value of annual report increases the predictive power of four machine learning models (Logistic regression, Random Forest, XGBoost, and Support Vector Machine) for corporate bankruptcy prediction using U.S. firm observations from 1994 to 2018. We find that the overall prediction effectiveness of these four models (e.g. accuracy, F1-score, AUCs) significantly improves, especially true in the performance of XGBoost and Random Forest models. In addition, we find that annual report text-based communicative value variables significantly reduce models' Type II error and keep the Type I error at a relatively small level, especially for the short-term bankruptcy forecast. The results reveal that annual report text-based communicative value effectively mitigates the model misidentification of a non-bankrupt firm as a bankrupt firm. Our results also suggest that annual report text-based communicative value is helpful for bank's corporate loan underwriting decisions. Finally, our findings still hold when considering different testing periods and random state settings.

Keywords: Annual report text-based communicative value, Bankruptcy prediction, Machine learning, Credit risk, Incomplete information

1. Introduction

In recent years, the global market has experienced the subprime mortgage crisis, the European debt crisis, the U.S.-China trade war, and the COVID-19 pandemic, which have caused economic and social turmoil, and many companies have also gone bankrupt due to their poor financial health. For instance, the subprime mortgage crisis in 2008 induces a global economic recession, which damages many firms' financial health condition, leads them to the dilemma of financial distress, and thus increases bank loan credit risk. Therefore, a firm's credit risk profile is very critical in the context of a poor economic environment. In addition, the subprime mortgage crisis also leads commercial banks to pay more attention to firm bankruptcy risk and therefore bankruptcy prediction turns to be an important issue both in academic literature and in practices. Since accurate bankruptcy prediction can not only reduce banks' loan default losses, but also improve their operating efficiency (namely reduce the phenomenon that misjudge clients' bankruptcy risk and thus refuse to grant credit) and ensure the stability of the financial system, effective bankruptcy prediction becomes a valued research issue in bank loan practices.

Among the previous studies related to bankruptcy prediction, most of them employ financial and accounting-related variables as input variables in the models, such as Beaver (1966), Altman (1968), and Ohlson (1980). Different from Beaver (1966) that only uses univariate analysis, Altman (1968) introduces five financial ratios (liquidity, profitability, productivity, leverage, and asset turnover)[1] as main input variables and employs discriminant analysis model as the model setting of bankruptcy prediction. Ohlson (1980) employs logistic regression model and introduces nine financial ratios to predict corporate bankruptcy. In recent years, with the advent of artificial intelligence and machine learning models, there have been new developed tools for bankruptcy prediction research to improve the accuracy of corporate bankruptcy prediction. Among them, most studies introduce financial factors as main input variables and employ various machine learning models to analyze and compare the model accuracy. (e.g. Barboza et al., 2017). The commonly used tools of machine learning models include Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM). Different from the previous studies applying machine learning models for corporate bankruptcy prediction, this study introduces a firm's annual report text-based communication value (hereafter denoted as *T_CV*) variables (Seebeck and Kaya, 2022) as main input variables in addition to the traditional financial ratio variables (Barboza et al., 2017) in various machine learning models and then conducts a comparative analysis of the predictive effectiveness of these models. This study

[1] In Altman (1968), the liquidity variable is defined as net working capital per unit asset; the profitability variable is defined as retained earnings per unit asset; the productivity variable is defined as earnings before interests and taxes per unit asset; the leverage variable is defined as the ratio of equity market value to total debts; and the asset turnover variable is defined as the ratio of net sales to total assets.

therefore aims to explore whether the firm's annual report *T_CV* variables empower the bankruptcy prediction models to capture more credit risk-related information and improve the effectiveness of bankruptcy predictions.

This study follows Seebeck and Kaya (2022) to employ readability and tone as main proxies of a firm's annual report *T_CV*, which can describe the degree of incomplete accounting information of the firm. According to structural-form credit risk models of Merton (1974) and Duffie and Lando (2001), asset value, asset value volatility, default threshold, and incomplete accounting information are four core determinants of firm credit risk. Hence, structural-form credit risk models serve as the theoretical foundation of using a firm's annual report *T_CV* variables as main input variables in bankruptcy prediction models with machine learning settings. Accordingly, this study employs annual report *T_CV* variables as main input variables in the machine learning model settings of Barboza et al. (2017) and uses the 11 financial characteristic variables selected by Barboza et al. (2017) as the benchmark model input variables. That is, under the setting of the aforementioned 11 financial characteristic variables, this study further adds the annual report *T_CV* variables to implement the machine learning models and explores whether the annual report text-based communication value enhances the bankruptcy prediction models to capture more credit risk-related information. Moreover, since the bankrupt firms are much lower than the non-bankrupt ones in our research sample data (namely data imbalance), this study not only focuses on the improvement on accuracy, but also the improvements on the Type I error and Type II error among the model performance indicators. This is mainly because the improvements on Type I error and Type II error can increase banks' decision-making effectiveness in loan business.

This study investigates whether the annual report text-based communicative value improves the effectiveness of bankruptcy prediction with machine learning models using American bankrupt and non-bankrupt firm data from 1994 to 2018. After the procedures of data preprocessing, feature selection, and data imbalance processing, this study implements the effectiveness analyses of bankruptcy predictions using four machine learning models, including Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM). We expect the model setting with additional including annual report *T_CV* variables into Barboza et al. (2017)'s 11 financial ratio (hereafter denoted as *Barboza_FR*) variables performs better than that only with *Barboza_FR* variables. In addition, to increase the robustness of the empirical results, this study firstly uses the data observations from 1994 to 2014 as the training group to predict corporate bankruptcy in the next one, two, three, and four years in a rolling year by year way. Moreover, we consider 100 groups of random state settings for each bankruptcy and employ the average of each model performance indictor's results in 100 random state settings as the final forecast result.

Empirical results of this study show that the overall prediction effectiveness of the model including the annual repot *T_CV* variables and *Barboza_FR* variables has a significant improvement compared to the model with only *Barboza_FR* variables. Among the performance indicators of our models, both accuracy rate and F1-score have substantial improvement, especially those of XGBoost and Random Forest models, implying that annual report text-based communication value variables can provide additional credit risk information beyond *Barboza_FR* variables for corporate bankruptcy prediction. In addition, compared to the model with *Barboza_FR* variables, this study also finds that the Type I error can be maintained at a relatively small level and the Type II error can be greatly reduced after adding annual repot *T_CV* variables into the bankruptcy prediction models. The economic implication therefore is that annual report text-based communicative value can effectively improve the situation where the model misjudges a non-bankrupt firm as a bankrupt firm. Hence, annual report text-based communicative value can increase the chances of the bank granting credit to normal customers, improve the efficiency of capital utilization, and effectively improve the performance of credit granting business. In addition, our main results still hold when considering different forecast periods (1 to 4 years in the future) and different random state settings. Finally, this study also finds that the annual report *T_CV* variables is more effective in predicting bankruptcy events in a relatively short-term period (such as one-year), consistent with the theoretical concepts of Duffie and Lando (2001) and Yu (2005).

It has to be noted that although the previous literature considers the characteristics of financial report text and combines the deep learning models to discuss bankruptcy prediction (e.g. Mai et al., 2019); however, the financial report text feature variable used in literature is mainly word frequency, which does not include all the annual report text-based communication value variables (readability and tones) mentioned in Seebeck and Kaya (2022). In addition, Mai et al. (2019) only consider the texts in the Management's Discussion and Analysis (MD&A) section of annual report rather than those in all sections of the annual report, such as the section of Notes to Consolidated Financial Statements (Chen and Tseng, 2021). Moreover, Mai et al. (2019) do not provide solid theoretical foundations and economic implications of introducing word frequency and the role of financial report texts on prediction effectiveness for different bankruptcy forecast periods (e.g. debt term structure). Meanwhile, Mai et al. (2019) mainly focus on model accuracy rather than Type I error and Type II error, and thus they could not provide further improvement suggestions about decision-making efficiency in bank credit loan practices.

Therefore, compared with Mai et al. (2019), the incremental contributions of this study are: (1). introducing the annual report *T_CV* variables (e.g. readability and tones) proposed by Seebeck and Kaya (2022) to corporate bankruptcy prediction model and providing the evidences that the annual report *T_CV* variables significantly improve the model performance indicators (e.g. accuracy, F1-score, Type II error, AUCs), especially for the short-term bankruptcy forecast; (2) providing solid theoretical foundations and economic implications for introducing the annual report *T_CV* variables into bankruptcy prediction models with short- and long-term forecasting periods (Duffie and Lando, 2001); (3) providing the evidences and economic implications that annual report *T_CV* variables can significantly improve the misjudgment of non-bankrupt companies as bankruptcy and can effectively enhance the performance of bank credit loan business. In summary, the results of this study not only fills the academic gap in bankruptcy prediction literature, but also provides suggestions for improving bank credit loan practices.

The remainder of this paper is organized as follows. Section 2 presents the literature reviews. Section 3 demonstrates the research methods, including data, research procedures, machine learning models, and evaluation indictors. Section 4 presents and analyzes empirical results. Finally, section 5 provides concluding remarks.

2. Literature Reviews

This section introduces the literature reviews and discussions in two subsections, focusing on the researches on (1) bankruptcy prediction and (2) the relationship between annual report *T_CV* variables (namely readability and tones) and credit risk. The first subsection introduces the existing bankruptcy prediction models (including statistical and machine learning models) and their employed financial ratio variables. The second subsection presents the theoretical foundations and economic implications for introducing the annual report *T_CV* variables into bankruptcy prediction models.

2.1. Bankruptcy prediction

Bankruptcy prediction has always been a critical issue in credit risk literature. Most of previous studies focus on employing financial ratio variables to predict whether firms face bankruptcy risks. As a pioneer in the field of bankruptcy prediction researches, Beaver (1966) employs univariate analysis and 14 financial ratios to predict the likelihood of corporate financial crisis. Altman (1968) employs Multiple Discriminant Analysis and 5 financial ratios to establish Z score as an early warning indicator for corporate bankruptcy, which can successfully predict 31 out of 33 bankrupt firms one year ahead. Since the prediction power of Altman's (1968) model is quite good, many subsequent studies related to bankruptcy prediction employ Altman's (1968) 5 financial variables as the fundamental input variables. Ohlson (1980) employs the logistic regression model and 9 financial ratio variables to predict corporate bankruptcy using America firm data from 1970 to 1976, covering

105 bankrupt firms and 2,058 non-bankrupt firms randomly selected. Ohlson (1980) finds 6 out of 9 financial ratio variables have significant impacts on bankruptcy predictions and believes that these financial ratio variables roughly determine the bankruptcy model's predictive power. Since most the previous studies related to bankruptcy prediction focus on financial ratio variables, non-financial variables may have the potential to help improve the predictive power of bankruptcy prediction models.

Among the previous studies related to bankruptcy prediction, financial accounting ratio variables generally serve as the main input variables of the prediction models, such as discriminant analysis (Altman, 1968) and logistic regression models (Ohlson, 1980). However, there exists improvement room on the accuracy of these traditional statistical-based bankruptcy prediction models (Begley et al.,1996). With the introduction of data science techniques, many related studies have gradually used machine learning or deep learning to predict corporate bankruptcy. Nanni and Lumini (2009) demonstrate that the predictive power of machine learning models is better than traditional statistical analysis methods. However, most of the previous studies on bankruptcy prediction focus on the improvement of the model accuracy by introducing new machine learning models with financial ratio variables and the comparison of the performance of machine learning models (e.g. Atiya, 2001; Shin et al., 2005; Kumar and Ravi, 2007; Tsai and Wu, 2008; Chen, 2011; Olson et al., 2012). The incremental improvement of model performance contributed by new input variables are rarely discussed (Jensen and Meckling, 1976; Barboza et al., 2017; Liang et al., 2017; Mai et al., 2019; Sun, 2020). Jensen and Meckling (1976) propose that non-financial factors should be considered when constructing financial crisis models. Barboza et al. (2017) introduce 6 new variables that significantly influence firm financial performance into bankruptcy prediction models in addition to the 5 financial ratio variables in Z score model (Altman, 1968). Liang et al. (2017) employ Altman's (1968) 5 financial ratio variables and add corporate governance variables as model input variables to compare the prediction performance before and after adding corporate governance variables. Moreover, Mai et al. (2019) and Sun (2020) introduce word frequency variables and high-order momentum risk information of equity market as additional model input variables to explore whether prediction performance has a significant improvement, respectively. Among the above mentioned studies, Liang et al. (2017), Mai et al. (2019) and Sun (2020) all employ non-financial variables with machine learning models on bankruptcy prediction.

In the bankruptcy prediction literature, most of them focus on the prediction power of machine learning models using financial variables. Few studies introduce the text-based communication value characteristics of annual report (readability and tone) as additional input variables in bankruptcy prediction with machine learning models. This study uses Barboza et al. (2017) as a benchmark model

(*Barboza_FR* variables, including 5 financial ratio variables and 6 financial performance change indicators) and additionally introduces annual report *T_CV* variables as new input variables to explore whether annual report *T_CV* variables improve the effectiveness of machine learning models for bankruptcy prediction. As for the selection of machine learning models, Barboza et al. (2017) demonstrate that the models such as SVM-RBF, Boosting, Bagging, and Random Forest perform well, so this study includes these models in the selection of machine learning model settings.

In addition, many previous studies on bankruptcy prediction with machine learning models focus on the improvement of accuracy. However, since there exists the serious data imbalance phenomenon in actual bankrupt and non-bankrupt firm data (namely the number of bankrupt firms is much lower than that of non-bankrupt firms), the improvement of accuracy cannot indeed ensure the improvement of Type I error and Type II error. Therefore, to provide a more precise reference for banks' decision-making considerations for credit loans, this study also focuses on whether Type I error and Type II error can be effectively improved.

2.2. The association between annual report text-based communication value and credit risk

Financial statements are one of the important ways for external investors to gain an in-depth understanding of a firm's operating conditions and development trends. Besides, the words used to explain the adopted accounting policy, the operating business, the management decisions, and the scale and the changes of accounting items in financial statements all have corresponding communicative value for external investors. Using audit report as an example, Seebeck and Kaya (2022) describe a firm's financial report communicative value by the following four measures, including readability, evaluative content, visual aids, and specificity. Therefore, this study follows the opinions of Seebeck and Kaya (2022) on the communicative value of annual reports, and defines annual report readability and tones (namely evaluative content) as the proxies of annual report text-based communicative value.

In the annual report readability literature, Li (2008) demonstrates that a firm's earnings performance is positively related to its annual report readability. That is, the annual reports of firms with larger earnings are easier to read than those of firms with poor profits (Subramanian et al. al., 1993). In addition, Lo et al. (2017) also find that a firm's earnings management level is negatively associated with the readability of the Management Discussions and Analyses (MD&A) section in the firm's annual report. Ajina et al. (2016) propose that the content of a firm's annual reports should include time, content and presentation, so as to improve the readability of annual reports. Ajina et al. (2016) also find that the firm with earnings decline has more incentives to increase the complexity of its annual reports in order to hide the fact of poor financial performance. The above studies have demonstrated that there is a certain relationship between a firm's financial performance and annual report readability. Since a firm's asset value is one of the main core factors affecting corporate credit risk (Merton, 1974), it is reasonable to expect that there is a certain correlation between annual report readability and firm bankruptcy risk.

According to the definition of linguistics, text readability can be regarded as the degree of comprehension of the text by the reader (Dale and Chall, 1949; McLaughlin, 1969; Klare, 1963). Therefore, readability can be defined as any of the characteristics of reading materials such as the ease of reading, legibility, and ease of understanding of the content in the text element (Klare, 1963). In previous related studies, most of them mainly discuss whether readability can help solve the gap between information users and information providers. For instance, the text information provided in financial reports may not be absolutely as a reference for external investors' investment decisions (namely low communicative value). Hence, annual report readability can also be regarded as one of the indicators of incomplete information, which increases the agency cost between the agent and the principal from the perspective of agency theory. Kothari (2000) demonstrates that external investors prefer to use higher-quality financial reporting information for decision-making in order to reduce the information asymmetry level and stock price fluctuations. Therefore, the higher annual report readability indicates the higher communicative value of the annual report and the lower implied agency cost.

In addition, information users have to spend more time and effort in extracting information when financial information disclosures are less readable. Bloomfied (2002) demonstrates that a firm's managers may take the opportunity to use confusing information to hide its poor performance when market investors are incomprehensible to public information. Previous studies document that the economic consequences of annual report readability include investment efficiency (Biddle et al, 2009), small investors' trading and investment behaviors (Miller, 2010; Lawrence, 2013), analysts' earnings forecast behaviors (Lehavy et al., 2011), and creditors' wealth effects (Bonsall and Miller, 2017; Chen and Tseng, 2021). Moreover, Guay et al. (2016) demonstrate that managers can also use voluntary disclosure to overcome the negative impact of complex financial statement content on the stock market. For the economic effect of annual report readability on creditors' wealth, Bonsall and Miller (2017) find that poor annual report readability leads to poorer credit rating scores (higher default risk), larger opinion dispersion of bond rating agencies, and higher cost of debt. Chen and Tseng (2021) also present that firms with higher readability of notes to consolidated financial statements have lower bond yield spread (namely lower default risk). Based on above discussions, we can therefore conclude that less readable disclosure in annual reports leads to lower credit rating and higher cost of debt (credit spreads), implying that less readable annual reports may signal higher likelihood of corporate bankruptcy.

In recent years, some studies introduce the text features of the annual report into the machine learning models for bankruptcy prediction as new input variables (e.g. Mai et al., 2019). However, Mai et al. (2019): (1) only use the word frequency features in the MD&A section of annual report rather than those in all sections of the annual report and do not consider the annual report text-based communicative value variables (Seebeck and Kaya, 2022); (2) do not clearly demonstrate the theoretical foundations and economic implications of annual report textual features on bankruptcy prediction, especially for short-term forecasting period; (3).do not provide further suggestions for improving the decision-making efficiency in bank credit loan practices. Different from the previous studies, this study introduces the annual report text-based communicative value variables (namely *T_CV* variables) into the machine learning models for bankruptcy predictions with short- and long-term forecasting periods. This study aims to compare and analyze the effectiveness of each machine learning model and further provides the policy implications of the annual report text-based communicative value characteristic variables in credit loan practices.

3. Research Methodology

This section introduces the research methodology of this study, including research procedure, data and research variables, data pre-processing, machine learning models, and confusion matrix. The details are introduced in the following.

3.1. Research procedure

The research procedures of this study are as follows: First, collecting bankrupt and non-bankrupt firm data from Compustat database, implementing data pre-processing (including removing data with null values, eliminating outliers, and data standardization), handling data imbalance, and using feature selection to identify important variables for the model. Next, in the process of data splitting, this study uses time as the basis for data splitting, and uses a long time period of sample data (e.g., more than 20 years) for training, and then tests the future data patterns and compares the prediction performance. Finally, this study employs confusion matrix and related model evaluation metrics (Accuracy, F1-score, Type I error, Type II error) to validate the prediction performance of the model.

3.2. Data and variables

This study employs all U.S. bankrupt and non-bankrupt firms from 1994 to 2018 as research sample and the related financial data used in this study are obtained from the Compustat database. In addition, the data of annual report text-based communicative value characteristics (i.e. readability and tones) are obtained from the SEC Analytics Suite database. Moreover, to identify whether a firm is bankrupt or not, this study employs Compustat DLRSN CODE to classify the firms as the bankrupt ones when their DLRSN CODE are 02 or 03. The data set is labeled as 1 for bankrupt firms and 0 for non-bankrupt firms, and finally 932 bankrupt firms and 40,507 non-bankrupt firms are included in the data set after data processing.

In terms of the use of financial input variables, this study follows Barboza et al. (2017) to employ 11 financial ratio variables as input variables in benchmark model setting. These 11 *Barboza_FR* variables include five financial variables that constitute Altman (1968)'s Z score and six variables of financial performance changes. The five Altman's (1968) Z score component variables cover the ratio of net working capital to total assets (NWC_TA), the ratio of retained earnings to total assets (RE_TA), return on assets (EBIT_TA), the ratio of equity market value to total debts (EMV_Debt), and the ratio of net sales to total assets (Sales_TA). The additional six variables of financial performance changes include the ratio of earnings before interests and taxes to net sales (namely EBIT margin, EBIT_Sales), the change rate of total assets (TA_growth), the change rate of net sales (Sales_growth), the change rate of the number of employee (EMP_growth), the change in return on equity (ROE_Chg), and the change in equity market-to-book value ratio (PB_Chg). The detailed definitions of the above 11 variables are shown in Table 1.

<Table 1> Variables Definitions: Barboza et al. (2017)

Variable	Formula
NWC_TA	$\frac{\text{Net working capital}}{\text{Total assets}}$
RE_TA	$\frac{\text{Retained earnings}}{\text{Total assets}}$
EBIT_TA	$\frac{\text{Earnings before interest and taxes}}{\text{Total assets}}$
EMV_Debt	$\frac{\text{Market value of share} * \text{number of shares}}{\text{Total debt}}$
Sales_TA	$\frac{\text{Sales}}{\text{Total assets}}$
EBIT_Sales	$\frac{\text{Earnings before interest and taxes}}{\text{Sales}}$
TA_growth	$\frac{\text{Total assets}_t - \text{Total assets}_{t-1}}{\text{Total assets}_{t-1}}$
Sales_growth	$\frac{\text{Sales}_t - \text{Sales}_{t-1}}{\text{Sales}_{t-1}}$

EMP_growth	$\frac{\text{Number of employee}_t - \text{Number of employee}_{t-1}}{\text{Number of employee}_{t-1}}$
ROE_Chg	$\text{ROE}_t - \text{ROE}_{t-1}$
PB_Chg	$\text{Price\_to\_Book}_t - \text{Price\_to\_Book}_{t-1}$

Note: Table 1 shows the definitions of financial variables employed in Barboza et al. (2017). The first five variables are mentioned by Altman (1968) and the rest six variables are based on Carton and Hofer (2006).

Regarding the annual report text-based communicative value (*T_CV*) characteristics variables, this study employs all readability and sentiment variables listed in the SEC Analytics Suite database. Then, this study employs Random Forest algorithm as feature selection tool, ranks the importance of the annual report *T_CV* variables, and identifies the top 25 variables that contribute to the model's prediction power (Please see Appendix A). Next, we exclude the variables with the coefficients of correlation greater than 0.7 and finally 11 annual report *T_CV* variables are retained. The detailed definitions of the selected *T_CV* variables are shown in Table 2.

<Table 2> Variables Definitions: Annual Report Text-based Communicative Value Variables

Variable	Definitions
FSIZE	The file size of 10-K document (unit: megabytes)
FK_Ease	Flesch Reading Ease Index of 10-K: 206.835 − 1.015(*number of words/number of sentences*) − 84.6(*number of syllables/ number of words*)
CL	Coleman Readability Index of 10-K: 5.88(*number of* characters/ *number of words*) − 29.6(*number of sentences/ number of words*) − 15.8
A_WPP	Average number of words per paragraph of 10-K: Number of words/ the number of paragraphs in the annual report
FT_NEG	Negative words proportion (Loughran-McDonald) in 10-K: The number of Loughran-McDonald Financial-Negative words in the annual report/ the total number of words in the annual report that occur in the master dictionary.
FT_MWeak_C	Modal weak word count (Loughran-McDonald) in 10-K: The number of Loughran-McDonald Financial-Modal-Weak words in the annual report.
FT_POS	Positive words proportion (Loughran-McDonald) in 10-K: The number of Loughran-McDonald Financial-Positive words in the annual report/ the total number of words in the annual report that occur in the master dictionary.
Harv_NEG	Negative words proportion (Harvard General Inquirer) in 10-K: The number of Harvard General Inquirer Negative words in the annual report.
FT_LITI	Litigious words proportion (Loughran-McDonald) in 10-K: The number of Loughran-McDonald Financial-Litigious words in the annual report/ the total number of words in the annual report that occur in the master dictionary.
FT_UNC	Uncertainty words proportion (Loughran-McDonald) in 10-K: The number of Loughran-McDonald Financial-Uncertainty words in the annual report/ the total number of words in the annual report that occur in the master dictionary.
FT_MStr	Modal Strong words proportion (Loughran-McDonald) in 10-K: The number of Loughran-McDonald Financial-Modal Strong words in the annual report/ the total number of words in the annual report that occur in the master dictionary.

Note: Table 2 shows the definitions of the selected annual report text-based communicative value (T_CV) variables based on Random Forest feature selection model. The original annual report T_CV variables shown in WRDS SEC Analytics Suite Database include 36 text-related variables. The definitions of the selected annual report T_CV variables are referenced from WRDS SEC Analytics Suite Database.

Table 3 presents summary statistics of bankruptcy indicator (*Brupt*), 11 *Barboza_FR* variables, and 11 annual report *T_CV* characteristics variables for the bankrupt firms, and non-bankrupt firms, full firm sample observations, respectively. The *Brupt* variable is a bankruptcy dummy variable that equals 1 if a firm has encountered bankruptcy and 0 otherwise. It has to be noted that the average of the *Brupt* variable is 0.022, implying that only 2.2% of the full sample firms are bankrupt. Hence, there exists a serious data imbalance concern for the research sample of corporate bankruptcy.

<Table 3> Summary Statistics of Major Variables

Panel A. Overall firm sample

	Obs.	Mean	Stdev	Min	Median	Max
NWC_TA	41439	0.193	1.539	-238.840	0.190	0.962
RE_TA	41439	-0.737	11.720	-1860.476	0.093	6.121
EBIT_TA	41439	-0.002	1.128	-165.699	0.066	1.745
EMV_Debt	41439	439.620	8847.059	0.001	4.352	750555.520
Sales_TA	41439	1.793	0.021	0.595	1.793	2.444
EBIT_Sales	41439	-4.128	192.266	-30175.700	0.064	394.474
TA_growth	41439	0.139	0.731	-0.989	0.045	57.133
Sales_growth	41439	1.040	88.222	-6.488	0.067	12739.000
EMP_growth	41439	0.103	1.460	-1.000	0.023	165.667
ROE_Chg	41439	-0.010	8.947	-658.630	-0.002	1117.200
PB_Chg	41439	-0.055	5.299	-79.993	0.000	79.993
FSIZE	41439	6.283	9.941	0.001	1.615	413.988
FK_Ease	41439	25.720	4.588	-187.265	25.484	52.930
CL	41439	22.356	0.808	18.459	22.281	39.423
A_WPP	41439	176.595	2626.008	11.252	68.882	209364.000
FT_NEG	41439	0.016	0.005	0.000	0.016	0.051
FT_MWeak_C	41439	224.676	175.097	0.000	193.000	3094.000
FT_POS	41439	0.008	0.002	0.000	0.008	0.027
Harv_NEG	41439	0.041	0.007	0.000	0.041	0.069
FT_LITI	41439	0.011	0.006	0.002	0.010	0.066
FT_UNC	41439	0.014	0.004	0.000	0.014	0.031
FT_MStr	41439	0.003	0.001	0.000	0.003	0.012
Brupt	41439	0.022	0.148	0.000	0.000	1.000

Panel B. Bankrupt firms

	Obs.	Mean	Stdev	Min	Median	Max
NWC_TA	932	0.1691	0.4031	-6.1570	0.1695	0.8500
RE_TA	932	-1.0469	4.7240	-126.1260	-0.1675	1.9420
EBIT_TA	932	-0.0867	0.3256	-3.3790	0.0140	1.0660
EMV_Debt	932	93.8632	757.4240	0.0020	1.8765	14156.0600
Sales_TA	932	1.7930	0.0000	1.7930	1.7930	1.7930
EBIT_Sales	932	-0.8260	5.1163	-115.0410	0.0100	0.4770
TA_growth	932	0.0841	0.5163	-0.9288	-0.0050	7.1603
Sales_growth	932	0.1781	0.8953	-0.9271	0.0353	19.9770
EMP_growth	932	0.2646	4.0941	-1.0000	-0.0084	93.4444
ROE_Chg	932	-0.2415	2.9694	-61.9610	-0.0328	22.2740
PB_Chg	932	-0.2266	5.9995	-79.9930	-0.0793	71.6187
FSIZE	932	3.0960	6.4071	0.0336	0.8844	74.5386
FK_Ease	932	27.5102	4.3388	11.1860	27.2241	43.3344
CL	932	22.2398	0.8247	20.4652	22.1155	26.1407
A_WPP	932	187.1657	1736.9876	19.7156	69.5759	31487.0000
FT_NEG	932	0.0157	0.0051	0.0025	0.0159	0.0306
FT_MWeak_C	932	182.8916	157.2439	2.0000	146.5000	879.0000
FT_POS	932	0.0081	0.0020	0.0007	0.0078	0.0181

Harv_NEG	932	0.0396	0.0068	0.0167	0.0399	0.0612
FT_LITI	932	0.0103	0.0051	0.0026	0.0092	0.0451
FT_UNC	932	0.0132	0.0036	0.0048	0.0134	0.0244
FT_MStr	932	0.0030	0.0011	0.0005	0.0028	0.0118

Panel C. Non-bankrupt firms

	Obs.	Mean	Stdev	Min	Median	Max
NWC_TA	40507	0.194	1.555	-238.840	0.191	0.962
RE_TA	40507	-0.730	11.832	-1860.476	0.096	6.121
EBIT_TA	40507	0.000	1.140	-165.699	0.067	1.745
EMV_Debt	40507	447.575	8947.366	0.001	4.406	750555.520
Sales_TA	40507	1.793	0.021	0.595	1.793	2.444
EBIT_Sales	40507	-4.204	194.463	-30175.700	0.066	394.474
TA_growth	40507	0.141	0.736	-0.989	0.046	57.133
Sales_growth	40507	1.060	89.231	-6.488	0.068	12739.000
EMP_growth	40507	0.100	1.339	-1.000	0.023	165.667
ROE_Chg	40507	-0.005	9.038	-658.630	-0.002	1117.200
PB_Chg	40507	-0.051	5.282	-79.993	0.000	79.993
FSIZE	40507	6.357	9.996	0.001	1.637	413.988
FK_Ease	40507	25.679	4.586	-187.265	25.448	52.930
CL	40507	22.358	0.807	18.459	22.285	39.423
A_WPP	40507	176.352	2642.960	11.252	68.872	209364.000
FT_NEG	40507	0.016	0.005	0.000	0.016	0.051
FT_MWeak_C	40507	225.637	175.371	0.000	194.000	3094.000
FT_POS	40507	0.008	0.002	0.000	0.008	0.027
Harv_NEG	40507	0.041	0.007	0.000	0.041	0.069
FT_LITI	40507	0.011	0.006	0.002	0.010	0.066
FT_UNC	40507	0.014	0.004	0.000	0.014	0.031
FT_MStr	40507	0.003	0.001	0.000	0.003	0.012

Note: Panels A, B, and C are the (pre-winsorized) descriptive statistics of the full sample, the bankrupt firm sample, and the non-bankrupt firm sample, respectively. The definitions of Barboza et al. (2017)'s financial variables and annual report T_CV variables can be referred to Table 1 and Table 2, respectively.

3.3. Data pre-processing

3.3.1. Null value and outlier processing

Since there are many missing values of accounting item in Compustat database, the employed financial ratio variables (*Barboza_FR*) have missing value problem. To avoid the distortions of the missing values of input variables on the machine learning models, we exclude the annual firm observations with any missing values of input variables. Therefore, the final sample observations are ensured to have input variables with non-null values.

In addition, to avoid the possible distortions of outliers on model prediction performance, this study winsorizes all input variables at upper and lower 1% levels.

3.3.2. Data standardization and data splitting

Since the value scales of different variables are not the same, to avoid the scale distortions on machine learning training process, this study employs standardization method to convert the value of each input variable into a standard constant assignment with a mean of 0 and a standard deviation of

1.

Before implementing machine learning models, this study splits the research sample data set into the training data set and the testing data set based on the criteria of time. The sample data before a certain year is used as the training data set and the sample data after a certain year is used as the testing data set. The training data set is used to build the machine learning models for bankruptcy prediction, and then the testing data set is used to verify the models' prediction results. In this study, the reasons why using time as the basis for splitting sample data are that (1) a firm's financial structure information varies from year to year; and (2) training the model with data in a continuous year interval allows the model easier to learn the data properties.

This study employs the period from 1994 to 2014 as the basic sample period of training data and the next year to the next four years as the sample period of testing data. It has to be noted that the study adopts a rolling adjustment for the sample period of the training data and that of the testing data year by year. That is, the training data from 1994 to 2014 (1994 to 2015; 1994 to 2016; 1994 to 2017) is used to predict corporate bankruptcy in 2015 (2016; 2017; 2018). As for the bankruptcy forecast for the next two years, the training data from 1994 to 2014 (1994 to 2015; 1994 to 2016) is used to predict corporate bankruptcy for the years 2015 to 2016 (2016 to 2017; 2017 to 2018). The corporate bankruptcy predictions for the next three years and the next four years follow a similar approach.

3.3.3. Imbalanced data processing

In bankruptcy prediction literature, the distribution of research sample data is usually imbalanced, meaning that there exists an imbalance between the sample size of bankrupt firms and non-bankrupt firms. Based on the Compustat's bankruptcy criteria, there are 40,507 sample observations of non-bankrupt firms and only 932 sample observations of bankrupt firms. To overcome the data imbalance limitation, this study employs two algorithms, EasyEnsemble and BalanceBaggingClassifier, as main methods to deal with data imbalance.

3.3.3.1 EasyEnsemble

The concept of EasyEnsemble is to perform random sampling k times in the majority category sample, and in each sampling, the same number of samples as the minority category sample are taken to generate k datasets, and then these data sets are trained k times to generate k different models, and the final result is obtained by a majority vote. This study employs EasyEnsemble method to generate 1000 subsets by randomly sampling 1000 times, and then generates 1000 sub-models for majority decision to obtain the final result. Therefore, EasyEnsemble is an integrated learning algorithm that combines Bagging and Adaboost.

3.3.3.2. BalanceBaggingClassifier

The BalanceBaggingClassifier is a data imbalance processing method that adds a base classifier to EasyEnsemble. A base classifier can be set using the parameter setting base_estimator, and the final classification results are obtained by combining multiple models.

3.3.4. Feature selection

In a dataset full of many variables, there are several important topics in feature selection process, such as: (1) how to select the variables that are effective for the model; (2) reducing the dimensionality and complexity of the model; and (3) removing those variables that are not useful for the model. This study employs the random forest algorithm to rank the importance of annual report *T_CV* variables (readability and tones). The principle of random forest algorithm for ranking the importance of variable features is simply to examine how much each feature contributes to each tree, then calculate the average value, and finally compare the contribution of different features to each other. The results of features selection in this study are shown in Figure 1.

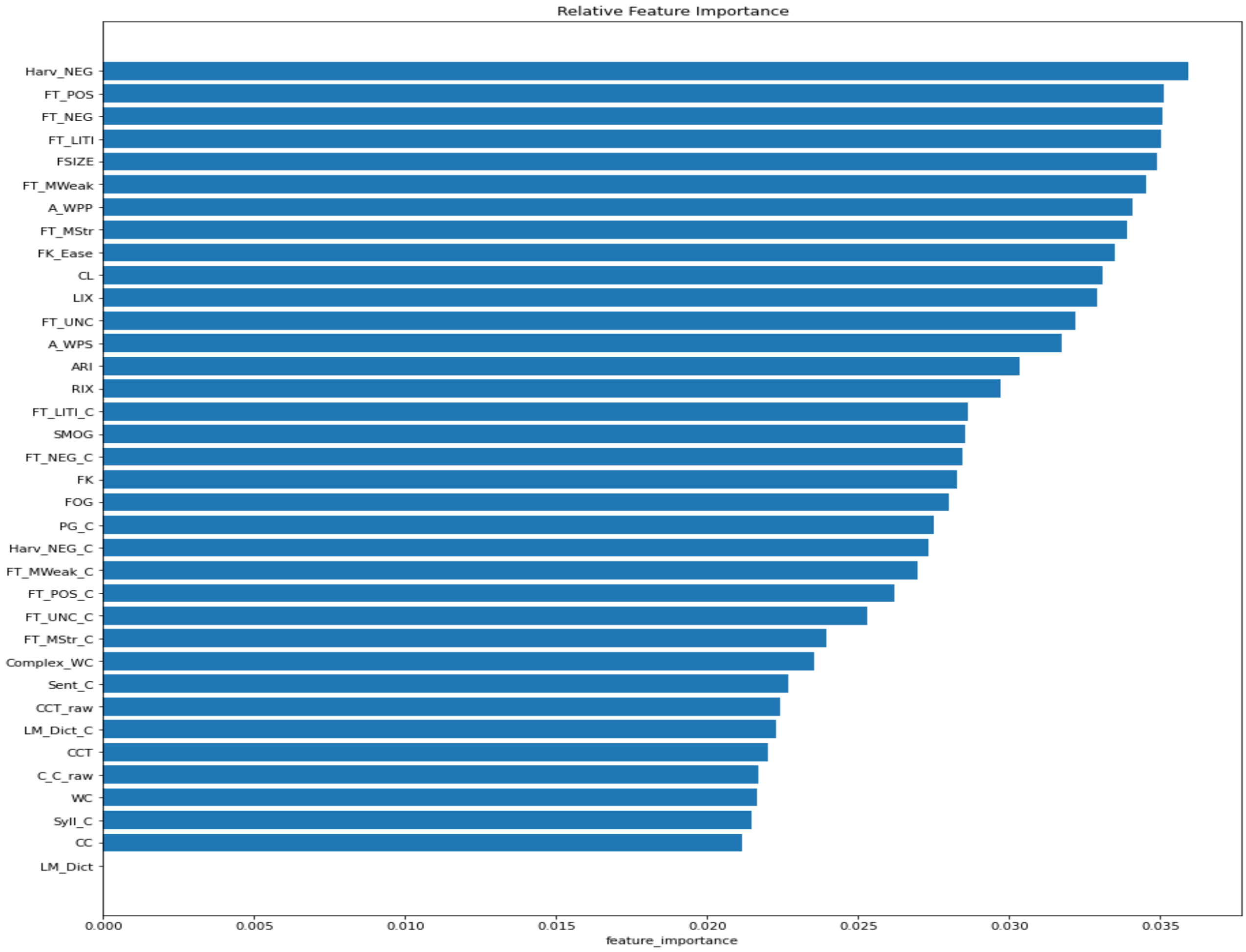

Figure 1 shows the importance ranking of the annual report *T_CV* variables based on random forest algorithm, and identifies the top 25 variables that contribute to the model's prediction power. Next, we exclude the variables with the coefficients of correlation greater than 0.7 and finally 11 annual report *T_CV* variables are retained.

<Figure 1> Relative Importance of Annual Report Text-based Communicative Value Variables

3.4. Machine learning models

Regarding the performance of machine learning models in bankruptcy prediction literature, Barboza et al. (2017) show that the top prediction performance models are Boosting, Random Forest, and Support Vector Machine (SVM), which the accuracy of these models are close to 90%. Hence, this study introduces these three machine learning models to implement the main analyses and discussions. Since XGBoost is a popular machine learning algorithm related to Boosting and perform well in many Kaggle's competitions, this study employs XGBoost as the proxied model of Boosting. In addition, this study also includes the Logistic Regression model, which is widely used to deal with classification problems in statistical-based models.

3.4.1. Logistic Regression

Logistic regression, based on maximum likelihood method, is a fundamental model of machine learning that is widely used to deal with classification problems. The target variable in logistic regression model is a binary category variable and a regression line is found to classify it correctly. This study sets a bankruptcy dummy variable that equals 1 if a firm is bankrupt and 0 if otherwise. The formula of logistic regression model can be shown as Eq. (1).

$$f(x)_\theta = \frac{1}{1+e^{-x}} \tag{1}$$

where θ is the weight value of each input variable and x are the input variables of the firm. Using the Sigmoid function, this study can estimate the probability value P under the corresponding input variable, which is also the firm's probability of bankruptcy, shown as Eq. (2).

$$\mathrm{P}(h(x)_\theta = 0{,}1|\theta_0, \dots., \theta_n) = \frac{1}{1+e^{-(\theta_0+\theta_1 x1+\cdots..+\theta_N x_N)}} \tag{2}$$

In addition, we can further employ the loss function to check the error rate of the model and find the optimal parameter θ by minimizing the error. Compared to other machine learning models, the logistic regression is easier to understand and can explain the generated results in an understandable way.

3.4.2. Support Vector Machine (SVM)

Support Vector Machine (SVM, Vapnik 1963) model, one of supervised machine learning models, is usually used to solve classification problems. The concept of SVM is to find a decision boundary (e.g. line or hyperplane) to maximize the margin between two categories so that the two categories can be perfectly separated. For the linearly indistinguishable data set, SVM-RBF (Radial

Based Function) is more applicable than SVM-Linear.[2] In previous studies related to bankruptcy prediction, the data distribution patterns of bankrupt firms and non-bankrupt firms are linearly indistinguishable, which leads the results of using SVM-RBF to be better than SVM-Linear (Min and Lee, 2005; Barboza, et al., 2017). Therefore, this study employs SVM-RBF to implement the model analyses.

3.4.3. XGBoost

Extreme Gradient Boosting (XGBoost, Chen and Guestrin, 2016) algorithm, one of machine learning models, has high training efficiency speed and excellent learning effect so that it is regarded as the main model in many machine learning related researches. XGBoost is one of the Boosting algorithms and the principle of the Boosting algorithm is to assemble many weak classifiers into a strong classifier. In the training process, the data error weight of the old classifier will be increased, and then the new classifier will continue to be trained, so that the new classifier and the subsequent training can learn the characteristics of the misclassified data and achieve improvement.

XGBoost is based on the extension and improvement of Gradient Boosted Decision Tree (GBDT) model, which is applied to solve the problem of supervised learning by aggregating many tree models into a strong classifier, and using gradient descent to minimize the residuals in the process. In addition, XGBoost also adds a regularization term as a penalty to prevent the model from overfitting. The employed tree model is a Classification and Regression Tree (CART) model, shown as Eq. (3).

$$\hat{y}_i = \sum_{n=1}^{N} f_n(x_i) \tag{3}$$

Where $\hat{y_i}$ indicates the prediction result of the model, N is the total number of decision trees, and f_n indicates the nth decision tree.

The objective function of XGBoost is composed of two parts, namely the loss function and the regularization term, shown as Eq. (4). For the loss function, it is used to measure the difference between the real result and the predicted result. The loss function contains $\hat{y}_i$ and $\hat{y}_i{}^{(p-1)}$, which represent the value of the current tree model ($f_p(x_i)$) and the predicted result of the previous tree, respectively. The purpose is to correct the residuals of each tree in the past and the residuals of the newly added trees. For the regularization term, it aims to solve the overfitting problem, which can be effectively prevented by adjusting the penalty value by controlling the hyperparameters γ and λ.

$$Obj^{(p)} = \sum_{i=1}^{k} l(y_i, \hat{y}_i{}^{(p-1)} + f_p(x_i)) + \Omega(f_p) \tag{4}$$

[2] The difference between SVM-RBF and SVM-Linear is that SVM-RBF uses Kernel Function to transform and map the data to higher dimensional space, so that the linearly indistinguishable data set can be classified in higher dimensional space using hyperplane to achieve linear differentiation.

$$\text{Where } \Omega(f_p) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \text{ ,}$$

In Eq. (4), k represents the number of samples in the training set, and p represents the number of trees constructed. T indicates the size of the tree, which means the number of leaf nodes, and w indicates the weight of the leaf nodes.

3.4.4. Random Forest

Random Forest algorithm (Breiman, 2001) is a model constructed by multiple decision trees, in which each constructs a decision tree and the final model result is decided by voting. That is, the main concept of random forest model is Bagging (Bootstrap Aggregation), which constructs a subset of data by repeated sampling, models and predicts them separately, and finally aggregates the results of all tree predictions to determine the classification result by a multi-decision approach. In the decision-making process, the decision tree calculates the amount of information (Entropy) at each node in each level of the tree. Then, the information gain (IG) is obtained by subtracting the weighted average information of the nodes from the classified information, and the feature with greater information gain is selected as the categorization basis. The calculation of information gain (IG) is shown as Eq. (5). In Eq. (5), Entropy(P) is the amount of information in the classified node P, and IG(B) is the weighted average amount of information before the classification is subtracted from the information after the classification by dividing the set P into k equal portions with feature B.

$$\text{Entropy(P)} = \ -\textstyle\sum_{i=1}^{x} p_i * \ Log^{p_i} \qquad (5)$$

$$\text{IG(B)} = Entropy(P) - \ \sum_{j=1}^{k} \frac{|P_j|}{|P|} * Entropy(P_j)$$

where P is a node in the set decision tree to make a decision; *x* is the number of categories in the set; *p* is the proportion of each category in the set.

The random forest model has the following features, such as: (1) it is more efficient than decision trees; (2) it is more efficient in dealing with missing values and outliers; (3) it is less prone to overfitting problems; (4) the model executes quickly and can make reasonable predictions without tuning hyperparameters.

3.5. Confusion Matrix

Confusion Matrix is a tool commonly used in machine learning to evaluate the classification and prediction results of models. The prediction results of Confusion Matrix can be classified into four categories: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). It has to be noted that Negative (Positive) means corporate bankruptcy (non-bankruptcy).[3] Among them, the evaluation variables used by the confusion matrix after calculation include Accuracy, Precision, Recall, F1-Score, Type I Error, and Type II Error. Among the above evaluation variables, the accuracy rate is defined as the percentage of correct predictions among all samples. The precision rate indicates how many of the samples predicted to be positive (negative) by the model are actually positive (negative) samples. The recall rate indicates how many positive (negative) samples the model is able to successfully predict from actually positive (negative) samples. The F1-Score is a weighted average of the precision rate and recall rate, shown as Eq. (6).

$$\mathrm{F1-Score} = 2 * \frac{Precision*Recall}{Precision+Recall} \tag{6}$$

The machine learning analyses in this research focus on model accuracy and misclassification. The misclassification scenarios are Type I error and Type II error, where Type I error means that a real bankrupt firm is predicted to be a non-bankrupt firm and Type II error means that a real non-bankrupt firm is predicted to be a bankrupt firm. This study expects that the machine learning model with the input of annual report *T_CV* variables can not only increase the accuracy rate but also improve both the Type I Error and Type II Error.

4. Empirical Analyses

This section presents the bankruptcy prediction results employing the above mentioned four machine learning models under different conditions, and further analyzes and discusses whether annual report T_CV variables enhance the effectiveness of bankruptcy prediction under the model based on Barboza et al. (2017). In addition to the model accuracy rate, this study specifically focuses on F1-score, Type I error, and Type II error. To provide stable model results, this study implements the prediction effectiveness analyses by employing 100 groups of random states for each machine learning model, and then obtains the prediction effectiveness by the average of each evaluation variable in 100 random states.

Table 4 shows the empirical results of bankruptcy prediction in the next year using the four machine learning models before and after adding the annual report *T_CV* variables. Panel A in Table 4 presents the results of bankruptcy prediction using machine learning models with *Barboza_FR* variables, and Panel B demonstrates the results with both *Barboza_FR* and annual report *T_CV*

[3] In this research, since Positive means non-bankruptcy and Negative means bankruptcy, True Positive means actual non-bankruptcy and the prediction result is also a non-bankruptcy sample; True Negative represents actual bankruptcy and the prediction result is also a bankruptcy sample; False Positive represents actual bankruptcy and the prediction result is a non-bankruptcy sample; False Negative represents actual non-bankruptcy and the prediction result is a bankruptcy sample.

variables. From the results of Panel A and B in Table 4, this study finds the prediction effectiveness of the four machine learning models is significantly improved after adding annual report *T_CV* variables using both BalancedBagging and EasyEnsemble algorithms. For example, under the Random Forest model with BalancedBagging algorithm and after adding annual report *T_CV* variables, the prediction accuracy increases from 79.80% to 91.84%, F1-score of the bankrupt firms (F1_1) increases from 6.03% to 13.68%, F1-score of the non-bankrupt firms (F1_0) increases from 88.68% to 95.71%, the Type II error decreases from 20.26% to 8.14%, and Type I error decreases from 9.23% to 7.75%.

In addition, among the four machine learning models with two data imbalance processing algorithms, the prediction accuracy has an average increase ranging from 10.5% to 16.01%; The F1-score of the bankrupt firms (F1_1) increases by 1.9% to 8.34%; the F1-score of the non-bankrupt firms (F1_0) increases by 6.02% to 10.29%; the Type II error decreases by 10.16% to 16.15%. The above model evaluation variables show that the annual report *T_CV* variables indeed improve the predictive power of the bankruptcy models. In particular, the Random Forest (RF) model with the data imbalance processing method of BalancedBagging has the best performance based on the criteria of Type I and Type II error (0.0775 and 0.0814).

Regarding the other three machine learning models, the XGBoost model performs relatively better than the SVM model and the Logistic regression model. The XGBoost model can greatly reduce the Type II error when the Type I error increases slightly, and the significant decrease in the Type II error can also be viewed as an important source of improving the overall accuracy of the bankruptcy prediction model. Therefore, we preliminarily conclude that adding the annual report *T_CV* variables as new input variables of bankruptcy prediction models leads to a great reduction on Type II error under a certain level of Type I error. That is, the annual report *T_CV* variables can greatly reduce the probability of misjudging a non-bankrupt firm as a bankrupt firm. The above results preliminarily provide the practical implications that the annual report *T_CV* variables can increase the bank's chances of granting credit loans to normal customers, improve the efficiency of bank funding utilization, and effectively improve the performance of credit loan operations.

<Table 4> Comparative Analyses of the Effectiveness of Bankruptcy Prediction Models (Forecasting period is the next year)

Panel A. Bankruptcy Prediction Using Machine Learning Models with Barboza et al. (2017)'s Financial Variables

	Accuracy	F1_1	F1_0	TP	TN	FN	FP	Type II error	Type I error
A.1.BalancedBagging									
Logistic	0.7504	0.0416	0.8565	1505.3450	11.3150	498.9050	4.1850	0.2494	0.2446
RF	0.7980	0.0603	0.8868	1599.7250	13.7125	404.5250	1.7875	0.2026	0.0923
SVM	0.7400	0.0499	0.8494	1481.4675	14.2650	522.7825	1.2350	0.2613	0.0553
XGBoost	0.8243	0.0709	0.9029	1652.6100	14.0200	351.6400	1.4800	0.1763	0.0671

A.2. EasyEnsemble									
Logistic	0.7437	0.0408	0.8521	1491.6150	11.4275	512.6350	4.0725	0.2563	0.2397
RF	0.6922	0.0435	0.8166	1384.3600	14.7550	619.8900	0.7450	0.3097	0.0340
SVM	0.6907	0.0448	0.8154	1381.0450	15.2050	623.2050	0.2950	0.3114	0.0148
XGBoost	0.7492	0.0536	0.8553	1501.6250	14.6350	502.6250	0.8650	0.2523	0.0368
Panel B. Bankruptcy Prediction Using Machine Learning Models with Barboza et al. (2017)'s Financial Variables and Annual Report Text-based Communicative Value variables									
	Accuracy	F1_1	F1_0	TP	TN	FN	FP	Type II error	Type I error
B.1. BalancedBagging									
Logistic	0.8792	0.0610	0.9354	1765.5450	8.6500	238.7050	6.8500	0.1184	0.4443
RF	**0.9184**	**0.1368**	**0.9571**	**1842.4025**	**13.7850**	**161.8475**	**1.7150**	**0.0814**	**0.0775**
SVM	0.8706	0.0750	0.9304	1745.1125	11.5525	259.1375	3.9475	0.1284	0.2578
XGBoost	**0.9293**	**0.1543**	**0.9631**	**1864.7425**	**13.6000**	**139.5075**	**1.9000**	**0.0703**	**0.0843**
B.2. EasyEnsemble									
Logistic	0.8740	0.0619	0.9324	1754.4700	9.1900	249.7800	6.3100	0.1239	0.4210
RF	**0.8523**	**0.0816**	**0.9195**	**1704.7900**	**14.3900**	**299.4600**	**1.1100**	**0.1482**	**0.0541**
SVM	0.8192	0.0638	0.8999	1640.5650	13.3025	363.6850	2.1975	0.1811	0.1399
XGBoost	**0.8931**	**0.1093**	**0.9431**	**1790.5100**	**14.0000**	**213.7400**	**1.5000**	**0.1070**	**0.0685**

Note: Table 4 show the results of prediction effectiveness whether a firm has a bankruptcy event in the next one year after adding the annual report text-based communicative value (T_CV) variables. The basic training period is from 1994 to 2014 and the training period is rolling adjusted year by year to predict corporate bankruptcy events in the next year. That is, the training period from 1994 to 2014 (1994 to 2015; 1994 to 2016; 1994 to 2017) is used to predict corporate bankruptcy in 2015 (2016; 2017; 2018). This study also performs 100 groups of random states for each bankruptcy prediction model with machine learning algorithm. The values in the table are the average of the rolling forecast results of each year under 100 groups of random states. In Table 4, the bankruptcy event is defined as 1 (called Negative), and the non-bankruptcy event is defined as 0 (called Positive). TP, TN, FN, and FP are the values of the confusion matrix, F1_1 represents the F1-score under the predicted bankruptcy event, and F1_0 represents the F1-score under the predicted non-bankruptcy event.

Table 5, 6, and 7 present the results of the model predictions of bankruptcy in the next two, three, and four years before and after adding the annual report *T_CV* variables, respectively. The results of Table 5, 6, and 7 show that the prediction accuracy of these four models is significantly improved using the data imbalance processing methods of BalancedBagging and EasyEnsemble, and the Type II error can be greatly reduced when the Type I error increases slightly. In addition, among these four machine learning models, we also find that the Random Forest (RF) and XGBoost models under BalancedBagging perform relatively better, consistent with the result of bankruptcy prediction in the next year (namely Table 4). Therefore, this study concludes that annual report *T_CV* variables can significantly reduce the Type II errors of predicting corporate bankruptcy events in the next one, two, three, and four years. That is, annual report *T_CV* variables are helpful for reducing the probability of misjudging non-bankrupt firms as bankrupt firms. This also verifies that the annual report text-based communicative value can increase the bank's chances of granting credit loans to normal customers, improve the efficiency of bank funding utilization, and effectively improve the performance of credit loan operations.

<Table 5> Comparative Analyses of the Effectiveness of Bankruptcy Prediction Models: (Forecasting period is the next two years)

Panel A. Bankruptcy Prediction Using Machine Learning Models with Barboza et al. (2017)'s Financial Variables

	Accuracy	F1_1	F1_0	TP	TN	FN	FP	Type II error	Type I error
A.1.BalancedBagging									
Logistic	0.7509	0.0435	0.8568	3075.5267	23.9500	1017.1400	8.7167	0.2490	0.2515
RF	0.7991	0.0643	0.8874	3268.2633	29.3067	824.4033	3.3600	0.2017	0.0891
SVM	0.7398	0.0524	0.8492	3024.2100	30.3533	1068.4567	2.3133	0.2617	0.0536
XGBoost	0.8263	0.0751	0.9042	3381.2133	29.7300	711.4533	2.9367	0.1743	0.0671
A.2. EasyEnsemble									
Logistic	0.7440	0.0427	0.8522	3046.8300	24.1333	1045.8367	8.5333	0.2560	0.2471
RF	0.6909	0.0456	0.8156	2820.4600	31.3033	1272.2067	1.3633	0.3112	0.0312
SVM	0.6910	0.0471	0.8156	2821.1200	32.2733	1271.5467	0.3933	0.3113	0.0084
XGBoost	0.7516	0.0559	0.8570	3074.2633	30.8500	1018.4033	1.8167	0.2498	0.0430

Panel B. Bankruptcy Prediction Using Machine Learning Models with Barboza et al. (2017)'s Financial Variables and Annual Report Text-based Communicative Value variables

	Accuracy	F1_1	F1_0	TP	TN	FN	FP	Type II error	Type I error
B.1. BalancedBagging									
Logistic	0.8772	0.0636	0.9343	3599.8133	17.9500	492.8533	14.7167	0.1203	0.4522
RF	**0.9184**	**0.1395**	**0.9572**	**3760.2567**	**28.1167**	**332.4100**	**4.5500**	**0.0812**	**0.1025**
SVM	0.8644	0.0814	0.9268	3538.9233	25.7500	553.7433	6.9167	0.1351	0.1923
XGBoost	**0.9279**	**0.1575**	**0.9623**	**3800.2067**	**28.5300**	**292.4600**	**4.1367**	**0.0717**	**0.0952**
B.2. EasyEnsemble									
Logistic	0.8715	0.0645	0.9310	3575.3767	19.1367	517.2900	13.5300	0.1262	0.4242
RF	**0.8445**	**0.0838**	**0.9150**	**3448.6100**	**30.8533**	**644.0567**	**1.8133**	**0.1563**	**0.0395**
SVM	0.8122	0.0672	0.8956	3321.2167	28.9000	771.4500	3.7667	0.1884	0.1015
XGBoost	**0.8890**	**0.1110**	**0.9408**	**3637.0400**	**29.6667**	**455.6267**	**3.0000**	**0.1112**	**0.0682**

Note: Table 5 show the results of prediction effectiveness whether a firm has a bankruptcy event in the next two years after adding the annual report text-based communicative value (T_CV) variables. In Table 5, the bankruptcy event is defined as 1 (called Negative), and the non-bankruptcy event is defined as 0 (called Positive).

<Table 6> Comparative Analyses of the Effectiveness of Bankruptcy Prediction Models: (Forecasting period is the next three years)

Panel A. Bankruptcy Prediction Using Machine Learning Models with Barboza et al. (2017)'s Financial Variables

	Accuracy	F1_1	F1_0	TP	TN	FN	FP	Type II error	Type I error
A.1.BalancedBagging									
Logistic	0.7504	0.0445	0.8565	4607.8200	35.9250	1531.1800	13.0750	0.2494	0.2648
RF	0.7973	0.0653	0.8863	4889.6200	43.8800	1249.3800	5.1200	0.2035	0.1008
SVM	0.7403	0.0537	0.8495	4535.5650	45.5300	1603.4350	3.4700	0.2612	0.0679
XGBoost	0.8258	0.0764	0.9038	5065.4800	44.5950	1073.5200	4.4050	0.1749	0.0850
A.2. EasyEnsemble									
Logistic	0.7432	0.0436	0.8517	4562.8650	36.2000	1576.1350	12.8000	0.2567	0.2592
RF	0.6885	0.0464	0.8139	4213.5900	46.9550	1925.4100	2.0450	0.3136	0.0395
SVM	0.6902	0.0481	0.8150	4222.4550	48.4100	1916.5450	0.5900	0.3122	0.0106
XGBoost	0.7504	0.0566	0.8562	4597.3600	46.2750	1541.6400	2.7250	0.2511	0.0546

Panel B. Bankruptcy Prediction Using Machine Learning Models with Barboza et al. (2017)'s Financial Variables and Annual Report Text-based Communicative Value Variables

	Accuracy	F1_1	F1_0	TP	TN	FN	FP	Type II error	Type I error
B.1. BalancedBagging									
Logistic	0.8780	0.0662	0.9347	5405.9550	26.8050	733.0450	22.1950	0.1194	0.4479
RF	**0.9161**	**0.1396**	**0.9559**	**5626.3700**	**42.1650**	**512.6300**	**6.8350**	**0.0835**	**0.1297**
SVM	0.8586	0.0815	0.9234	5274.2800	38.7800	864.7200	10.2200	0.1409	0.2004
XGBoost	**0.9248**	**0.1553**	**0.9606**	**5679.7150**	**42.7950**	**459.2850**	**6.2050**	**0.0748**	**0.1206**
B.2. EasyEnsemble									
Logistic	0.8723	0.0677	0.9315	5369.3950	28.7000	769.6050	20.3000	0.1254	0.4102
RF	**0.8341**	**0.0825**	**0.9088**	**5115.0300**	**46.2800**	**1023.9700**	**2.7200**	**0.1668**	**0.0499**
SVM	0.8044	0.0669	0.8907	4934.0500	43.3400	1204.9500	5.6600	0.1963	0.1100
XGBoost	**0.8816**	**0.1082**	**0.9366**	**5411.0650**	**44.5000**	**727.9350**	**4.5000**	**0.1186**	**0.0863**

Note: Table 6 show the results of prediction effectiveness whether a firm has a bankruptcy event in the next three years after adding the annual report text-based communicative value (T_CV) variables. In Table 6, the bankruptcy event is defined as 1 (called Negative), and the non-bankruptcy event is defined as 0 (called Positive).

<Table 7> Comparative Analyses of the Effectiveness of Bankruptcy Prediction Models: (Forecasting period is the next four years)

Panel A. Bankruptcy Prediction Using Machine Learning Models with Barboza et al. (2017)'s Financial Variables

	Accuracy	F1_1	F1_0	TP	TN	FN	FP	Type II error	Type I error
A.1.BalancedBagging									
Logistic	0.7489	0.0428	0.8555	6004.8100	45.3700	2012.1900	16.6300	0.2510	0.2682
RF	0.7911	0.0611	0.8825	6336.0900	54.9200	1680.9100	7.0800	0.2097	0.1142
SVM	0.7382	0.0512	0.8482	5907.2100	57.0600	2109.7900	4.9400	0.2632	0.0796
XGBoost	0.8207	0.0710	0.9008	6575.1800	55.3200	1441.8200	6.6800	0.1798	0.1077
A.2. EasyEnsemble									
Logistic	0.7414	0.0419	0.8506	5944.4400	45.6900	2072.5600	16.3100	0.2585	0.2631
RF	0.6825	0.0439	0.8096	5454.6200	58.9200	2562.3800	3.0800	0.3196	0.0497
SVM	0.6877	0.0460	0.8133	5494.9700	60.8200	2522.0300	1.1800	0.3146	0.0190
XGBoost	0.7438	0.0536	0.8519	5950.7000	58.5500	2066.3000	3.4500	0.2577	0.0556

Panel B. Bankruptcy Prediction Using Machine Learning Models with Barboza et al. (2017)'s Financial Variables and Annual Report Text-based Communicative Value Variables

	Accuracy	F1_1	F1_0	TP	TN	FN	FP	Type II error	Type I error
B.1. BalancedBagging									
Logistic	0.8805	0.0638	0.9362	7080.8400	32.9200	936.1600	29.0800	0.1168	0.4690
RF	**0.9097**	**0.1225**	**0.9524**	**7298.3400**	**50.9100**	**718.6600**	**11.0900**	**0.0896**	**0.1788**
SVM	0.8553	0.0739	0.9215	6863.0000	46.6700	1154.0000	15.3300	0.1439	0.2473
XGBoost	**0.9184**	**0.1383**	**0.9572**	**7366.8500**	**52.8900**	**650.1500**	**9.1100**	**0.0811**	**0.1469**
B.2. EasyEnsemble									
Logistic	0.8743	0.0652	0.9326	7028.2400	35.4300	988.7600	26.5700	0.1233	0.4286
RF	0.8248	0.0746	0.9032	6606.7300	57.0000	1410.2700	5.0000	0.1759	0.0806
SVM	0.7999	0.0623	0.8880	6409.1300	53.6900	1607.8700	8.3100	0.2006	0.1340
XGBoost	0.8712	0.0957	0.9307	6983.6900	55.0000	1033.3100	7.0000	0.1289	0.1129

Note: Table 7 show the results of prediction effectiveness whether a firm has a bankruptcy event in the next four years after adding the annual report text-based communicative value (T_CV) variables. In Table 7, the bankruptcy event is defined as 1 (called Negative), and the non-bankruptcy event is defined as 0 (called Positive).

To provide more robust evidences regarding the effectiveness of bankruptcy prediction models after adding the annual report *T_CV* variables, this study implements the mean differences tests of Type I and Type II errors using the two samples of the machine learning models with 100 random states before and after adding the annual report *T_CV* variables. The results of mean difference tests are shown in Table 8. The results of the mean difference tests show that the annual report *T_CV* variables can significantly reduce the Type II error under each machine learning model. However, the significant reduction of Type I error after adding annual report *T_CV* variables is only achieved under the Random Forest (RF) model with the data imbalance processing method of BalancedBagging. In addition, this study also finds that the annual report *T_CV* variables is more effective in predicting bankruptcy events in a relatively short-term period (such as one-year). The result is consistent with Duffie and Lando (2001) and Yu (2005) that incomplete accounting information have higher explanatory power for short-term corporate bond yield spread (namely credit risk).

<Table 8> The Difference Tests Analyses in the Effectiveness of Bankruptcy Prediction Models: Financial Variables v.s. Annual Report Text-Based Communicative Value Variables

Panel A. Forecasting Period is the Next One Year

Model	Measure	Mean_FIN	Mean_FIN_Text	Difference	T statistic	p-value
Panel A. BalancedBagging						
Logistic	Type II error	0.2494	0.1184	-0.1311	-204.0056	0.0000
	Type I error	0.2446	0.4443	0.1997	43.7406	0.0000
XGBoost	Type II error	0.1763	0.0703	-0.1060	-275.7449	0.0000
	Type I error	0.0671	0.0843	0.0172	12.8168	0.0000
RF	**Type II error**	**0.2026**	**0.0814**	**-0.1212**	**-376.1864**	**0.0000**
	Type I error	**0.0923**	**0.0775**	**-0.0147**	**-6.0860**	**0.0000**
SVM	Type II error	0.2613	0.1284	-0.1328	-178.8267	0.0000
	Type I error	0.0553	0.2578	0.2025	35.8194	0.0000
Panel B. EasyEnsemble						
Logistic	Type II error	0.2563	0.1239	-0.1324	-210.4173	0.0000
	Type I error	0.2397	0.4210	0.1812	36.8337	0.0000
XGBoost	Type II error	0.2523	0.1070	-0.1453	-146.2057	0.0000
	Type I error	0.0368	0.0685	0.0318	11.1385	0.0000
RF	Type II error	0.3097	0.1482	-0.1615	-149.0470	0.0000
	Type I error	0.0340	0.0541	0.0201	8.1872	0.0000
SVM	Type II error	0.3114	0.1811	-0.1303	-203.2852	0.0000
	Type I error	0.0147	0.1399	0.1252	37.1334	0.0000

Panel B. Forecasting Period is the Next Two Years

Model	Measure	Mean_FIN	Mean_FIN_Text	Difference	T statistic	p-value
Panel A. BalancedBagging						
Logistic	Type II error	0.2490	0.1203	-0.1287	-286.0788	0.0000
	Type I error	0.2515	0.4522	0.2007	63.0707	0.0000
XGBoost	Type II error	0.1743	0.0717	-0.1026	-416.0764	0.0000
	Type I error	0.0671	0.0952	0.0281	17.4741	0.0000
RF	**Type II error**	**0.2017**	**0.0812**	**-0.1205**	**-479.1368**	**0.0000**
	Type I error	**0.0891**	**0.1025**	**0.0135**	**3.3576**	**0.0009**
SVM	Type II error	0.2617	0.1351	-0.1266	-198.4545	0.0000
	Type I error	0.0536	0.1923	0.1388	51.6882	0.0000
Panel B. EasyEnsemble						
Logistic	Type II error	0.2560	0.1262	-0.1297	-298.9519	0.0000
	Type I error	0.2471	0.4242	0.1771	46.6441	0.0000
XGBoost	Type II error	0.2498	0.1112	-0.1387	-164.4934	0.0000
	Type I error	0.0430	0.0682	0.0252	11.9675	0.0000
RF	**Type II error**	**0.3112**	**0.1563**	**-0.1549**	**-151.5042**	**0.0000**
	Type I error	**0.0312**	**0.0395**	**0.0083**	**5.6022**	**0.0000**
SVM	Type II error	0.3113	0.1884	-0.1229	-227.1910	0.0000
	Type I error	0.0084	0.1015	0.0931	53.7216	0.0000

Panel C. Forecasting Period is the Next Three Years

Model	Measure	Mean_FIN	Mean_FIN_Text	Difference	T statistic	p-value
Panel A. BalancedBagging						
Logistic	Type II error	0.2494	0.1194	-0.1300	-534.1665	0.0000
	Type I error	0.2648	0.4479	0.1831	69.0729	0.0000
XGBoost	Type II error	0.1749	0.0748	-0.1001	-484.9510	0.0000
	Type I error	0.0850	0.1206	0.0356	25.2341	0.0000
RF	**Type II error**	**0.2035**	**0.0835**	**-0.1200**	**-578.1173**	**0.0000**
	Type I error	**0.1008**	**0.1297**	**0.0289**	**8.7853**	**0.0000**
SVM	Type II error	0.2612	0.1409	-0.1203	-459.4567	0.0000
	Type I error	0.0679	0.2004	0.1325	46.6205	0.0000
Panel B. EasyEnsemble						
Logistic	Type II error	0.2567	0.1254	-0.1314	-591.3470	0.0000
	Type I error	0.2592	0.4102	0.1510	92.8990	0.0000
XGBoost	Type II error	0.2511	0.1186	-0.1326	-278.2461	0.0000
	Type I error	0.0546	0.0863	0.0317	13.7138	0.0000
RF	**Type II error**	**0.3136**	**0.1668**	**-0.1468**	**-248.8690**	**0.0000**
	Type I error	**0.0395**	**0.0499**	**0.0104**	**5.6290**	**0.0000**
SVM	Type II error	0.3122	0.1963	-0.1159	-529.9995	0.0000
	Type I error	0.0106	0.1100	0.0995	46.7640	0.0000

Panel D. Forecasting Period is the Next Four Years

Model	Measure	Mean_FIN	Mean_FIN_Text	Difference	T statistic	p-value
Panel A. BalancedBagging						
Logistic	Type II error	0.2510	0.1168	-0.1342	-485.8132	0.0000
	Type I error	0.2682	0.4690	0.2008	70.9693	0.0000
XGBoost	Type II error	0.1798	0.0811	-0.0987	-370.1796	0.0000
	Type I error	0.1077	0.1469	0.0392	18.8483	0.0000
RF	**Type II error**	**0.2097**	**0.0896**	**-0.1200**	**-472.6698**	**0.0000**
	Type I error	**0.1142**	**0.1788**	**0.0647**	**43.3461**	**0.0000**
SVM	Type II error	0.2632	0.1439	-0.1192	-449.6852	0.0000
	Type I error	0.0796	0.2473	0.1676	80.7668	0.0000
Panel B. EasyEnsemble						
Logistic	Type II error	0.2585	0.1233	-0.1352	-690.1156	0.0000
	Type I error	0.2631	0.4286	0.1655	92.8187	0.0000
XGBoost	Type II error	0.2577	0.1289	-0.1289	-478.4950	0.0000
	Type I error	0.0556	0.1129	0.0573	58.3682	0.0000
RF	**Type II error**	**0.3196**	**0.1759**	**-0.1437**	**-440.6724**	**0.0000**
	Type I error	**0.0497**	**0.0806**	**0.0309**	**62.4426**	**0.0000**
SVM	Type II error	0.3146	0.2006	-0.1140	-577.1438	0.0000
	Type I error	0.0190	0.1340	0.1150	80.5156	0.0000

Note: The values of Mean_FIN (Mean_FIN_Text) are the average values of Type I error and Type II error of the rolling forecast results of each year under 100 groups of random states with Barboza et al. (2017)'s financial variables (Barboza et al. (2017)'s financial variables and annual report T_CV variables) as model input variables. Difference stands for Mean_FIN_Text minus Mean_FIN. The T-test statistic represents the difference test of the Type I error and Type II error of the rolling forecast results of each year in the above two groups of input variables under 100 random states.

Moreover, this study also presents the AUCs (Area Under the ROC Curve) of these bankruptcy prediction models using the machine learning algorithms of Random Forest (RF) and XGBoost, shown as Table 9. The results show that the bankruptcy prediction models perform better in future one, two, three, and four years after adding the annual report *T_CV* variables into these models. Meanwhile, the AUCs of the bankruptcy prediction models are higher in a shorter time period (e.g. one year) compared with other longer time periods (e.g. two, three, and four years). The results are robust for our conclusions that (1) the annual report text-based communicative value variables can further improve the effectiveness of bankruptcy prediction in addition to Barboza et al. (2017)'s financial variables; (2) the annual report *T_CV* variables is more effective in predicting bankruptcy events in a relatively short-term period, consistent with Duffie and Lando (2001) and Yu (2005) that incomplete accounting information is more pronounced for the short-term credit risk.

<Table 9> The AUCs Comparisons of Bankruptcy Prediction Models for Future Different Periods: Financial Variables v.s. Annual Report Text-Based Communicative Value Variables

Model	AUC in future 1 year	AUC in future 2 years	AUC in future 3 years	AUC in future 4 years
Panel A. 2015 (Beginning Prediction Time Point)				
XGBoost&Easy Ensemble				
FIN	0.9172	0.9080	0.9034	0.8973
FIN&T_CV	0.9559	0.9481	0.9400	0.9274
XGBoost&BalancedBagging				
FIN	0.9131	0.8874	0.8835	0.8586

FIN&T_CV	0.9418	0.9325	0.9191	0.8874
RF&Easy Ensemble				
FIN	0.9121	0.8924	0.8800	0.8793
FIN&T_CV	0.9480	0.9448	0.9321	0.9167
RF&BalancedBagging				
FIN	0.9165	0.9016	0.8983	0.8759
FIN&T_CV	0.9475	0.9414	0.9349	0.9125
Panel B. 2016 (Beginning Prediction Time Point)				
XGBoost&Easy Ensemble				
FIN	0.9197	0.9118	0.9077	
FIN&T_CV	0.9699	0.9599	0.9524	
XGBoost&BalancedBagging				
FIN	0.9226	0.9032	0.9128	
FIN&T_CV	0.9652	0.9555	0.9522	
RF&Easy Ensemble				
FIN	0.9146	0.8988	0.8816	
FIN&T_CV	0.9669	0.9609	0.9481	
RF&BalancedBagging				
FIN	0.9298	0.9133	0.9233	
FIN&T_CV	0.9666	0.9599	0.9595	

Note: In Table 9, FIN and T_CV present Barboza et al. (2017)'s 11 financial variables and our added annual report text-based communicative value (T_CV) variables, respectively.

5. Conclusions

The main purpose of this study is to explore whether the annual report text-based communicative value increases the predictive power of the bankruptcy prediction models with machine learning algorithm and captures more signals. In the recent years, although it is widely discussed in academic literature for bankruptcy predictions using various machine learning models; however, most of them only focus on financial variables rather than non-financial variables, which leads to a huge room for further improvement on the adoption of input variables. Our results show that adding the annual report *T_CV* variables not only significantly improve the effectiveness of the bankruptcy prediction models but also significantly improve the F1 score compared to Barboza et al. (2017) model. In addition, we also find that the annual report *T_CV* variables significantly reduce the probability of misjudging non-bankrupt firms as bankrupt ones (namely Type II error) under a certain level of Type I error. The above results suggest that the annual report text-based communicative value can help a bank's decision-making of granting credit loans and improve its funding utilization efficiency in practices. In addition, the model with the annual report *T_CV* variables have such a significant improvement in predictive power compared to Barboza et al. (2017) model, which also verifies the problems caused by the less readable financial information disclosure in Bonsall and Miller (2017), such as poorer credit rating scores (higher default risk), more divergent opinions among bond rating agencies, and higher debt costs. Therefore, we can conclude that the annual report text-based communicative value indeed has a significant influence on corporate bankruptcy prediction.

The main contributions of this study include: (1) introducing the concept of the annual report text-based communicative value into bankruptcy prediction models with machine learning algorithms; (2) providing the solid theoretical foundations and economic implications of the annual report text-based communicative value on bankruptcy prediction models; (3) discovering that the annual report text-based communicative value is beneficial for the effectiveness of short-term bankruptcy prediction and significantly reducing the probability of misjudging non-bankrupt firms as bankrupt ones; (4) providing the practical suggestions for banks to improve their funding utilization and operating efficiency. Finally, in addition to the annual report T_CV variables, we also suggest that future research can try to introduce other non-financial variables, or use different feature selection methods and machine learning algorithms, especially for other boosting and bagging methods.

References

Ajina, A., Laouiti, M., & Msolli, B. (2016). Guiding through the Fog: Does annual report readability reveal earnings management?. *Research in International Business and Finance, 38,* 509-516.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance,* 23(4), 589-609.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, 12(4), 929-935.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405-417.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71–111.

Begley, J., Ming, J. & Watts, S. (1996). Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models. *Review of Accounting Studies* 1, 267–284.

Biddle, G. C., Hilary, G., & Verdi, R. S. (2009). How does financial reporting quality relate to investment efficiency? *Journal of Accounting and Economics*, 48(2–3), 112–131.

Bloomfield, R. J. (2002). The'incomplete revelation hypothesis' and financial reporting. *Accounting Horizons* 16(3), 233-243.

Bloomfield, R. J., & Fischer, P. E. (2011). Disagreement and the cost of capital. *Journal of Accounting Research*, 49(1), 41–68.

Bonsall IV, S. B., & Miller, B. P. (2017). The impact of narrative disclosure readability on bond ratings and the cost of debt. *Review of Accounting Studies* 22 (2), 608-643.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Brockman, P., & Turtle, H. J. (2003). A barrier option framework for corporate security valuation. *Journal of Financial Economics*, *67*(3), 511-529.

Carton, R., & Hofer, C. (2006). Measuring organizational performance . Edward Elgar Publishing .

Chen, M. Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications*, *62*(12), 4514-4524.

Chen, T.K., & Tseng, Y. (2021). Readability of notes to consolidated financial statements and corporate bond yield Spread. *European Accounting Review* 30(1), 83-113

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: *Instructions. Educational Research Bulletin*, 37-54.

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19-26.

Duffie, D., & Lando, D. (2001). Term structures of credit spreads with incomplete accounting information. *Econometrica*, 69(3), 633-664.

Guay, W., Samuels, D., & Taylor, D. (2016). Guiding through the fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics*, *62*(2-3), 234-269.

Jensen, M.C. & Meckling (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3, 305-360.

Klare, G. R. (1963). Measurement of readability. University of Iowa Press, Ames, IA.

Kothari, S. P. (2000, June). The role of financial reporting in reducing financial risks in the market. In *Conference Series-Federal Reserve Bank of Boston* (Vol. 44, pp. 89-102). Federal Reserve Bank of Boston; 1998.

Kothari, S. P., Li, X., & Short, J. E. (2009). 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*, 84(5), 1639-1670.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European Journal of Operational Research*, *180*(1), 1-28.

Lehavy, R., Li, F., & Merkley, K. (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review*, *86*(3), 1087-1115.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*(2-3), 221-247.

Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics* 63(1), 1-25.

Liang, D., Lu, C. C., Tsai, C. F., & Shih, G. A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2), 561-572.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

Loughran, T., & McDonald, B. (2014a). Measuring readability in financial disclosures. *Journal of Finance*, 69(4), 1643–1671.

Loughran, T., & McDonald, B. (2014b). Regulation and financial disclosure: The impact of plain English. *Journal of Regulatory Economics*, 45(1), 94–113.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research* 274, 743–758

McLaughlin, G.H. (1969). SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8), 639-646.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2), 449-470.

Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *Journal of Finance*, 42(3), 483–510.

Min, J. H., & Lee, Y. C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems With Applications*, *28*(4), 603-614.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems With Applications*, *36*(2), 3028-3033.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109-131.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464-473.

Shin, K., Lee, T. S., & Kim, H. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1), 127-135.

Subramanian, R., Insley, R. G., & Blackwell, R. D. (1993). Performance and readability: A comparison of annual reports of profitable and unprofitable corporations. *The Journal of Business Communication (1973)*, *30*(1), 49-61.

Sun, Y.K., (2020). Bankruptcy prediction effectiveness analyses based on machine learning models: New evidences from higher order moment risks of equity market information. Master thesis, National Yang Ming Chiao Tung University.

Seebeck, A. & Kaya, D. (2022). The Power of Words: An Empirical Analysis of the Communicative Value of Extended Auditor Reports. *European Accounting Review*, forthcoming.

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems With Applications*, *34*(4), 2639-2649.

Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 774-780.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2 (3), 408-421.

Yu, F. (2005). Accounting transparency and the term structure of credit spreads. *Journal of Financial Economics* 75 (1): 53-84.