

“Using textual analysis in bankruptcy prediction: Evidence from Indian firms under IBC”

AUTHORS

Vandana Gupta 

Aditya Banerjee 

ARTICLE INFO

Vandana Gupta and Aditya Banerjee (2023). Using textual analysis in bankruptcy prediction: Evidence from Indian firms under IBC. *Investment Management and Financial Innovations*, 20(3), 22-34.
doi:[10.21511/imfi.20\(3\).2023.03](https://doi.org/10.21511/imfi.20(3).2023.03)

DOI

[http://dx.doi.org/10.21511/imfi.20\(3\).2023.03](http://dx.doi.org/10.21511/imfi.20(3).2023.03)

RELEASED ON

Wednesday, 05 July 2023

RECEIVED ON

Tuesday, 25 April 2023

ACCEPTED ON

Wednesday, 28 June 2023

LICENSE



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

JOURNAL

"Investment Management and Financial Innovations"

ISSN PRINT

1810-4967

ISSN ONLINE

1812-9358

PUBLISHER

LLC “Consulting Publishing Company “Business Perspectives”

FOUNDER

LLC “Consulting Publishing Company “Business Perspectives”



NUMBER OF REFERENCES

55



NUMBER OF FIGURES

0



NUMBER OF TABLES

5

© The author(s) 2023. This publication is an open access article.



BUSINESS PERSPECTIVES



LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine
www.businessperspectives.org

Received on: 25th of April, 2023

Accepted on: 28th of June, 2023

Published on: 5th of July, 2023

© Vandana Gupta, Aditya Banerjee,
2023

Vandana Gupta, Dr, Professor (Finance & Accounting), FORE School of Management, New Delhi, India. (Corresponding author)

Aditya Banerjee, Dr, Assistant Professor (Finance & Accounting), FORE School of Management, New Delhi, India.



This is an Open Access article, distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Conflict of interest statement:

Author(s) reported no conflict of interest

Vandana Gupta (India), Aditya Banerjee (India)

USING TEXTUAL ANALYSIS IN BANKRUPTCY PREDICTION: EVIDENCE FROM INDIAN FIRMS UNDER IBC

Abstract

Identifying and managing credit risk is vital for all lending institutions. Historically, credit risk is assessed using financial data from published financial statements. However, research indicates that the ability to detect financial hardship may be improved by textual analysis of firms' disclosed records. This study aims to establish an association between themes and words from Management Discussion and Analysis (MDA) reports of firms and corporate failures. The study took a sample of 57 Indian listed firms declared bankrupt under the Insolvency and Bankruptcy Code (IBC) along with a matched sample of 55 solvent firms (matched by industry and size) for the period of FY2011–2019. The first part of analysis identifies negative words from the published reports and compares them with the negative words of the Loughran-McDonald dictionary. Then a thematic analysis is done to identify the key themes from the MDA reports and the significant themes are validated with their corresponding financial ratios in the third step using a panel logistic regression. Word analysis results show that IBC firms have significantly greater negative tone (2.21 percent) as against 1.30 percent of solvent firms. Thematic analysis results show that manageability, activity and performance are significant themes for predicting financial distress. Financial variables such as ownership pattern, promoters' shares pledged, return on capital employed, asset utilization are some of the ratios in sync with the key themes. The study recommends that lenders and other stakeholders should look beyond financial statements which may be 'window dressed' by firms to qualitative disclosures in annual reports which may forewarn against impending financial distress.

Keywords

bankruptcy, word analysis, themes, disclosures, ratios, panel logit

JEL Classification

G33, M40, M41

INTRODUCTION

Credit risk arises from the possibility of loss when the borrower is unwilling or unable to repay the lender in full, leading to an economic loss to the lending bank. Banks may not have the necessary information to assess a borrower's default probability, which can lead to financial losses for the banks and, subsequently, cause a systemic crisis. Identifying red flags in borrowers' credit behavior can help banks mitigate credit risk. A firm's credit risk impacts all its stakeholders: the investors, lenders, and management; therefore, assessing and managing credit risk is essential.

Usually, credit risk is modeled with either linear discriminant analysis (LDA) or logistic regression (LR). However, these models require published financial data from financial reports as input. Bankruptcy prediction based on published financial reports is challenging due to the low frequency (quarterly or annual) of published information. Due to this, default prediction in the short term is difficult. Although accounting-based and market-based models continue to be applied

for credit risk and default prediction, financial data can be subject to ‘creative accounting’ practices or window dressing, as was evident in Enron and WorldCom cases and Satyam Computers in India (Gandhi, 2019).

Textual analysis of firms’ published documents has gained prominence in recent times. Textual analysis refers to techniques that are applied to elicit information from relevant texts. The information collected from texts can be used for research, analysis, and business intelligence (Loughran & McDonald, 2016). In the published annual reports of listed Indian companies, Management Discussion and Analysis (MDA) is a statement where firm executives comment on the past performance of their company, address the issues of compliance and risk, and present the outlook and proposed plan of action of a firm. In conjunction with financial statements, MDA provides valuable insights that could provide enhanced information about a firm’s credit risk and can be valuable input in credit risk models (Nguyen & Huynh, 2020). Therefore, textual analysis of specific sections of published financial documents has been adopted in this study as a key methodology for predicting financial distress.

1. LITERATURE REVIEW

The pioneers in developing accounting models were Beaver (1966, 1968) and Altman (1968). Beaver (1966) applied a univariate statistical analysis to predict corporate failure, while Altman (1968) applied the multiple discriminant analysis (MDA) and proposed the Z-Score model that could distinguish between distressed and solvent companies with financial ratios as input data. Other models included logistic regression by Ohlson (1980), and Zavgren (1988). Zmijewski (1984) developed a model by applying Probit Analysis, while Shumway (2001) pointed to the merits of applying panel data in bankruptcy prediction. Several models of bankruptcy prediction have been developed using machine learning (Yeh & Lien, 2009; Gepp et al., 2010; Barboza et al., 2017; Khemakhem & Boujelbene, 2018; Shrivastava et al., 2020).

The application of textual analysis techniques in finance and accounting is a novel and emerging field. However, given the volume of documents in financial disclosure, there is significant scope for extracting critical information from the published text. These disclosures can complement the information present in published financial statements (Abrahamson & Amir, 1996). The tone of the documents published in a disclosure process can reveal some underlying facts about a firm that may not be evident in its financial statements (Kearney & Liu, 2014; Amani & Fadlalla, 2017). The style and content of the disclosures in these reports often contains more relevant information than can be obtained from financial statements alone. Thus

textual-based predictors can serve as potential signals of financial distress among firms.

One of the first studies that applied textual analysis in finance was by Kohut and Segars (1992), which discriminated between companies with good and bad financial performance based on president’s letters. Swales (1988) applied thematic analysis and demonstrated that factors like growth, anticipated profits and losses, and management optimism could help differentiate between financially strong and weak firms. Abrahamson and Amir (1996), in their study, explained the importance of analyzing the annual report text for investors. Bryan (1997) and others advocate for businesses’ MD&A disclosures can help evaluate their short-term prospects.

The research by Smith and Taffler (1992) and Tennyson et al. (1990) use content analysis to examine the relationship between firms’ disclosures and bankruptcy. Ingram and Frazier (1980, 1983) conducted a descriptive study on the content of firms’ environmental disclosures for three different industries, and found a relationship between narrative disclosures and firm performance. Frazier et al. (1984) explored using the WORDS package (computer software) for analyzing accounting disclosures. His study found that the content of the documents is related to firms’ market performance. The management teams of surviving and bankrupt enterprises’ various responses to demand crises were examined using content analysis in the D’Aveni and MacMillan (1990) study. Previts et al. (1994) in their study revealed

that when making investment recommendations, analysts take into account both financial and non-financial information from the MD&A and the president's letter. According to Zhang et al. (2010), the dictionary-based textual analysis is related to the 'bag of words' since texts are considered unsorted sets of words. Four distinct word lists are popular in accounting and finance research. These are, The Henry (2008) list, the Harvard General Inquirer (GI), Diction, and the Loughran & McDonald (2011) list. The main weakness of the Henry list is that it was used originally for examining earnings press releases. A significant portion of the material might not be evaluated because of the Henry list's word count limitations. The Harvard dictionary provides a software-based mapping function for text files based on several dictionary classes using an algorithm to classify words in each category. The Harvard dictionary has been widely used for analyzing press news, corporate disclosures, and initial prospectuses due to its earlier availability. Another dictionary used is 'Diction,' that includes data on the overall word count, character count, average word length, number of distinct word types, counts of special characters, and high-frequency word counts. Harvard Dictionary and Diction did not have any specific vocabulary for financial applications. As a result, terms that are considered unfavorable in a general sense may have a different meaning in a financial sense. Li (2010b) found that tone classifications based on conventional dictionaries were not accurate enough.

Loughran and McDonald (2011) pioneered textual analysis by building a financial dictionary with six lists of words that offered more accuracy than the traditional Harvard Dictionary. The dictionary was used only to analyze financial communication, and it contained six lists of words to represent six different types of sentiment in that context (negative, positive, uncertainty, litigious, strong modal, and weak modal). The works were extended in 2014 and 2016 (Loughran & McDonald, 2014, 2016).

The role of the dictionary in financial research was extended by Engelberg et al. (2012) who revealed that competent information-processing short sellers use publicly available news as signals for their trades. A new readability metric for financial reporting was developed by Bonsall et al. (2017).

Their study demonstrated that firms can reduce their cost of debt, and achieve better credit ratings if their filings have better readability. Other similar research works were extended by Haralambie (2016) who examined the issues related to credit risk management considered by commercial banks when analyzing corporate clients. Hu et al. (2018) analyzed whether annual reports' readability affects credit default swap (CDS) market participants and how CDS affects firms' credit risk evaluation. Donovan et al. (2018) evaluated if information relevant to credit risk can be obtained by applying supervised machine learning to conference calls.

To the best knowledge, Gandhi et al. (2019) and Mai et al. (2019) are the two closest works that utilize textual data to predict firm default. Gandhi et al. (2019) use sentiment words to analyze the financial hardship of US banks, and their results show that more negative sentiment words in the reports increase the likelihood that the bank will be delisted owing to later financial distress. A deep learning model is used by Mai et al. (2019) to investigate the increase in prediction power brought about by word embedding methods. Smith and Taffler (2000) applied word frequency analysis which they termed "form orientated", and themes analysis, termed "meaning orientated".

The significance of theme analysis, one of the most popular research methodologies in qualitative research, was highlighted in works by Braun and Clarke (2006). The key objective of thematic analysis is to establish different themes, assign headings for the same from individual studies, and present them coherently (Thorne et al., 2004). Research questions and concepts are developed and explored using the key concepts and claims embodied by themes (Lián & Fayolle, 2015). A mixed-method approach can be adopted for the validation of textual analysis. Several works on mixed methods in finance have been done, including those by Creswell (1999), Östlund (2011), and Shorten and Smith (2017). All these authors have advocated that numbers and words together are essential to convey the results.

This study implements textual analysis through a mixed-method approach to identify the credit risk of Indian firms. To the best of the authors' knowl-

edge, such an approach has not yet explored in the context of Indian firms. This study can be broadly divided into three parts. The first part analyzes the difference in negative words in MDA reports among insolvent and healthy firms using the Loughran and McDonald (2016) dictionary. The study then proceeds to thematic analysis, where keywords from MDA reports are divided among existing themes (Smith & Taffler, 2000; Tate et al., 2010) and a new theme validated by professionals from the credit rating industry and the head of credit of an Indian bank. Thus, the validation by two industry experts covered both sides of the spectrum: the lender and the rating industry. The objective of the thematic analysis is to study whether MDA reports of insolvent firms classified in themes can indicate the possibility of bankruptcy in firms. In the third part, a quantitative analysis of the panel data of the financials of sample firms is carried out to determine the significant variables and thereby validate the themes analyzed in the second part.

To the best of the authors' knowledge, the association between textual disclosures (word analysis and thematic analysis) and corporate bankruptcy has not been done in the Indian context. Prior works are extended in testing if an association between firm failure and key 'words' and 'themes' is there for firms in India that are bankrupt under IBC (Insolvency and Bankruptcy Code). The findings from textual analysis are validated with quantitative analysis to see if the variables can map with the significant themes.

This study aims to establish an association between themes and words from Management Discussion and Analysis (MDA) reports of firms and corporate failures.

Based on the extant literature discussed above, three hypotheses are framed. In this context, the first hypothesis is:

H₁: Financial distress is more strongly correlated with the annual report's use of more negative words. Therefore, the proportions of negative words are statistically the same for IBC and solvent firms.

The second hypothesis is related to thematic analysis:

H₂: Themes generated from MDA reports of firms can discriminate between bankrupt and solvent firms.

The third hypothesis is related to the quantitative analysis used to validate our selected themes:

H₃: Financial variables corresponding to the significant themes can distinguish between IBC and solvent firms.

2. METHOD

2.1. Dataset and sample

The financial data and MDA reports of sample companies are collected from the Centre for Monitoring Indian Economy (CMIE), Prowess. 57 companies are identified that filed for bankruptcy in the financial years (FY) 2016–17 to 2018–19, classified under the IBC (Insolvency and Bankruptcy Code). IBC was introduced in November 2016, so the firms were included from FY 2017 onwards. The period from 2020 to 2021 is excluded due to the suspension of IBC proceedings for several companies by the Indian Government as a relief during the COVID-19 period¹.

55 solvent companies are identified from the same sectors as IBC companies that are rated 'AAA' and 'A' (highest and high safety, respectively) by India's leading credit rating agencies. The solvent companies are selected from the same industry if their average size (log of total assets) is within one standard deviation (on either side) of the size of IBC companies during the sample period. The MDA reports are collected and analyzed for all companies for two years before their respective filing dates. Since complete financial statements and MDAs were not readily available for most IBC firms beyond two years before filing, it presented a limitation in terms of data. Thus two years of MDA reports are taken for all firms in the sample. Financial data is collected for each company from 2010–11 to 2015–16 from the CMIE Prowess database. The financial data thus obtained was in an unbalanced panel format.

¹ Please see <https://ibclaw.in/suspension-of-ibc-during-covid-19-a-mere-mask-while-what-is-needed-is-a-vaccine-by-anvit-seemansh/?print-posts=print> for a discussion on this issue.

2.2. Mixed method approach

The mixed method approach is used in applied research and uses quantitative and qualitative methods. Researchers have advocated some theories and rationale for using this approach. Dewasiri et al. (2018) debate that confirming the results by two separate techniques increases the completeness, rationality, and soundness of findings than a single approach. Hasse-Biber (2010) and Greene et al. (1989) emphasize that the motivation for the mixed method approach can increase the validity while minimizing bias, thereby allowing analysis from different perspectives and expanding the overall rigor of the study.

This study uses mixed methods following the opinion of Jick (1979) and Creswell (2003), and looks for convergence across qualitative and quantitative methods within social science research on credit risk. A mixed-method approach is used in mapping the key financial indicators with the significant themes generated to validate the themes. The validation approach consists of a quantitative analysis using financial ratios mapped to the most significant themes identified from the selected themes under study.

2.3. Textual analysis: Word and thematic analysis

The initial analysis in the study aims to understand if IBC and comparable solvent firms differ concerning their tone in the MDA document. A two-sample test of proportion is used to compare the proportions of negative words in IBC and solvent firms. Considering the better contextual relevance of the Loughran and McDonald's (2011) dictionary to the financial statements, the same has been adopted for the initial analysis. The word analysis using the Loughran and McDonald method is applied to test the first research hypothesis formed earlier (H_1).

The Loughran and McDonald's (2011) dictionary can differentiate positive and negative tones but cannot explain the factors that could have led to the tone of the financial document. Therefore, thematic analysis is adopted as the second step, which can shed light on the factors that might be used to identify possible insolvency beforehand.

Thematic analysis is used to test the second hypothesis (H_2). The thematic analysis begins with auto-coding themes from MDA reports in the software. The significant themes are then identified using thematic analysis by the NVIVO software.

To test this hypothesis, logistic regression is run. The variable of interest (P_i) is a binary variable for i companies where $P_i = 1$ for companies that filed for bankruptcy and $P_i = 0$ for solvent companies. The explanatory variables (X_i) are the average theme scores for each firm obtained from the documents for the pre-IBC filing years. The scores are generated as the ratios of each theme's weight to the summation of theme coding references. Since single scores are obtained for each firm for each variable, the data being in a cross-section format. The logit model (L_i) used in this study can be expressed in a general form below:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \sum_{k=1}^6 \beta_k X_{k,i} + u_i, \quad (1)$$

$$P_i = \frac{1}{1 + e^{\beta_0 + \sum_{k=1}^6 \beta_k X_{k,i} + u_i}}, \quad (2)$$

where β_0 is the intercept of the equation, while β_k represents the slope coefficients for each of the six themes, u_i is the error term, and k is the six thematic variables mentioned earlier.

2.4. Quantitative analysis

The previous section's thematic analysis aims to identify the most important themes from the MDA reports that can differentiate between bankrupt and solvent companies. Assuming τ themes out of the six under study are relevant ($\tau \leq 6$), the next step is to determine the financial variables corresponding to those themes. With the financial variables corresponding to the relevant themes, it is possible to validate them and minimize any bias, as discussed earlier. To identify financial variables, prior research is explored where financial variables are classified into themes that are conceptually similar to those chosen in this study.

Since financial data is collected in an unbalanced panel format, panel logistic regression method is applied. With τ significant themes, X_τ variables

can be generated in the logistic regression model (G_i). The logistic model can be expressed as below:

$$G_{i,t} = \ln \left(\frac{Q_{i,t}}{1-Q_{i,t}} \right) = \delta_0 + \sum_{n=1}^{\tau} \delta_n X_{n,i,t} + \vartheta_{i,t}, \quad (3)$$

$$Q_{i,t} = \frac{1}{1 + e^{\delta_0 + \sum_{n=1}^{\tau} \delta_n X_{n,i,t} + \vartheta_{i,t}}}, \quad (4)$$

where $Q_{i,t}$ is a binary variable for i companies, where $Q_i = 1$ for companies that filed for bankruptcy and $Q_i = 0$ for solvent companies. δ_0 is the intercept of the equation while δ_n represents the slope coefficients for each of the significant themes, $\vartheta_{i,t}$ is the error term, and t represents the years in the panel.

3. RESULTS

3.1. Findings from word analysis

In the first step of the analysis, NVIVO 12 is run for word frequency analysis. The words generated from the MDA reports for both sets of firms are compared with the negative word list from Loughran and McDonald's (2011) dictionary. The results indicate that whereas there were 2.21 percent of bad words for the bankrupt firm, there were only 1.30 percent of negative words for solvent firms. Results from the two-sample test of proportion are available in Table 1. The test statistic value is greater than the z-statistic at the 5 percent level. Thus, the first hypothesis (H_1) is confirmed that more negative words from corporate disclosures are indeed associated with financial distress in companies.

Table 1. Result from the test of proportions

Source: Authors' computations.

| Word analysis | Firms under IBC | Solvent firms |
|-------------------------------|-----------------|---------------|
| Negative word count | 6984 | 5719 |
| Total Word Count | 315737 | 439130 |
| Percentage | 2.21% | 1.30% |
| Z-statistic (two-sample test) | 30.308 | |
| p-value | 0.000 | |

Note: This table shows the results from a pooled two-sample test of proportion. The proportion of negative words common to the Loughran-McDonald dictionary for each set of firms (IBC and solvent) is compared. The null hypothesis is that the proportion of negative words in IBC firms is greater than those in solvent firms with equivalent characteristics.

3.2. Findings from thematic analysis

The initial analysis results in the MDA contents being classified into 12 broad themes. Rather than relying entirely on the themes generated from the software, the themes are reclassified based on those available in the literature and validated by experts from rating agencies. Three categories for categorizing content qualities were established by Osgood et al. (1957): evaluative (positive/negative), potency (strong/weak), and activity (active/passive). From an intertemporal perspective, Houghton (1988) adds a fourth dimension to accounting, along with evaluative (beneficial/unfavorable), potency (tangible/intangible), activity (dynamic/static), and manageability (expected/unexpected). The research study also refers to Smith and Taffler's (2000) classification of the Chairman's comments of corporations into four categories: evaluative (beneficial/adverse), potency (tangible/intangible), activity (dynamic), and manageability (static/expected/unexpected). The critical sub-themes under these broad themes are classified based on the judgment and after discussion with eminent executives in the risk management domain. These themes shed light on the sentiments expressed in the qualitative disclosures of annual reports, indicating any financial distress.

The themes evaluative, activity, and management quality themes are maintained based on prior works on thematic analysis as stated above. A further dimension of outlook, performance, and strategic risks is added. The theme 'outlook' and 'strategic risks' is validated by industry experts and from prior research. 'Performance' theme is identified on the assumption that good/bad performance is always stated as part of MDA reporting and has also been classified as a relevant theme in prior similar research works. While strategic risks theme considers keywords as 'risks' and 'concern', outlook helps get an insight into prospects of 'investment,' 'projects,' 'infrastructure,' and 'developments' (Hanley & Hoberg, 2017).

The keywords are extracted for each theme before the computation of scores. A description of the number of keywords found for each theme and the weight of each theme is provided in Table 2.

Table 2. Themes and their weights

| Theme | References | Weight | Major Keywords |
|-----------------|------------|--------|--|
| Activity | 9470 | 0.14 | Business, demand, exports, imports, operations |
| Evaluative | 10078 | 0.15 | Effective, efficient, improvement, increase, changes |
| Management | 20478 | 0.30 | Control, internal, initiatives, governance, disclosures, value |
| Performance | 19399 | 0.28 | Growth, interest cost, profits, financial performance |
| Strategic risks | 4809 | 0.06 | Risks, concerns, going concern, operational risk, financial risk |
| Outlook | 4917 | 0.07 | Development, investment, infrastructure, projects |
| Total | 69151 | 1.00 | |

Note: This table shows the six themes adopted in this study and the weight assigned to each theme.

Equation 1 estimates the logistic model's parameters based on the average theme score for each company for each variable. The results of Equation 1 are presented in Table 3. The results show that the coefficients of only three themes are significant: activity, management quality, and performance. The coefficient of all the three significant themes are negative thereby implying inverse relationship between financial distress and themes.

Table 3. Results from logistic regression with theme scores

| Variables in the Equation | Coefficients (β_j) | Std. Errors |
|---------------------------|----------------------------|-------------|
| Constant (β_0) | 1.185*** | 0.317 |
| Activity Score | -0.441*** | 0.168 |
| Evaluative Score | -0.356 | 0.232 |
| Management Score | -0.181*** | 0.057 |
| Performance Score | 0.228*** | 0.069 |
| Strategic Risks Score | 0.321 | 0.717 |
| Outlook score | -0.242 | 0.352 |
| LR-Statistic (Chi-square) | 40.99*** | |

Note: This table shows the results of logistic regression in Equation 1. The dependent variable is binary which takes the value of 1 for IBC firms and 0 for solvent firms. The notation '***' refers to significance at the 1% level, '**' refers to significance at the 5% level, and '*' refers to significance at the 10% level.

The next step in this process is to identify the financial variables that correspond to these themes and verify the robustness of these themes with financial information.

3.3. Results from quantitative analysis

Results from the previous section show three significant themes: activity, management quality, and performance. The financial performance of a company can be assessed using a combination of financial ratios based on activity, management capability and performance. This research study

has identified 17 ratios under the broad classification stated above. The criteria for choosing ratios are those that: (i) They have been identified in the literature as default measuring indicators; (ii) Previous empirical works have applied these ratios in predicting insolvency; and (iii) They are easy to compute with data from exiting financial databases; (iv) The multi-collinearity among the variables is not significant (variance inflation factor below 5). The size (log of total assets) and age of firms are controlled. A description of the variables and the variance inflation factor (VIF for multi-collinearity) is provided in Table 4.

Table 5 contains the panel logistic regression's findings. Results show that 9 out of 17 variables are significant. Thus, two activity ratios (Gross fixed asset utilization ratio and Export/sales), four management quality ratios (Market capitalization/Debt, Debt/equity, Promoters' shareholding, and Promoters' shares pledged), and two performance ratios (Price/Book Value, and return on total assets) are statistically significant. Both control variables are also significant. Based on the results, we can say that our third hypothesis (H_3) is partially validated.

Activity ratios demonstrate how dynamic a company's operations are. Better operational efficiency is important for delivering quality goods or services to customers leading to a more profitable firm with higher credit ratings and thus credit-worthiness (Smith & Taffler, 2000). Performance of a firm has a direct bearing on its operational and financial risk as is evident by the profitability ratios (EPS, P/BV, PAT/Capital Employed and ROTA). The more profitable a company is, the better is its ability to service debt and thereby reduce its financial risk. Management quality is the most significant theme as is reflected in the ownership pattern, the quality of shares pledged, the market

Table 4. List of financial variables

| Variable | Meaning | Classification |
|--|--|----------------|
| Current ratio (CR) | Current Assets/Current Liabilities | Activity |
| Gross fixed asset utilization ratio | Turnover/Gross Fixed Assets | Activity |
| Export/sales | Percent of export sales/Total sales | Activity |
| Market capitalization /Debt | Market Capitalisation /outstanding debt | Management |
| Debt/equity (D/E) | Total borrowings/Shareholders' Funds | Management |
| Contingent Liabilities / Net Worth | Contingent liabilities/net worth | Management |
| Cash Flow from Operations /debt (CFO/Debt) | Cash Flow from Operations/ Debt | Management |
| Promoters' shareholding (%) | Promoters shares/Total shares | Management |
| Promoters' shares pledged (%) | Promoters share pledges/Total Promoters shares | Management |
| Earnings Per Share (EPS) | Net Profit/No. of equity shares | Performance |
| Price/Book Value (P/BV) | Market price per share/book value per share | Performance |
| PAT/capital employed | Net Profit/ Long term debt+ shareholders' funds | Performance |
| Finished goods cycle | Cost of Production/ Finished Goods Inventory*365 | Performance |
| Return on total assets (ROTA) | PBIT/Total assets | Performance |
| Net profit margin (NPM) | Net Profit/Sales *100 | Performance |
| Size | Natural log total assets | Control |
| Age | From the year of incorporation | Control |

Table 5. Panel logistic regression results

| Variable | Classification | Coefficients (δ_{ij}) | Std. Errors |
|--|----------------|--------------------------------|-------------|
| Constant (δ_0) | | 21.6835*** | 7.8376 |
| Current ratio (CR) | Activity | -1.2953 | 2.1253 |
| Gross fixed asset utilization ratio | Activity | 0.1341*** | 0.05177 |
| Export/sales | Activity | -0.1643** | 0.08104 |
| Market capitalization /Debt | Management | -0.2830** | 0.1186 |
| Debt/equity (D/E) | Management | 4.7373*** | 1.5572 |
| Contingent Liabilities / Net Worth | Management | 0.0021 | 0.0121 |
| Cash Flow from Operations /debt (CFO/Debt) | Management | 0.1655 | 1.4674 |
| Promoters' shareholding (%) | Management | -0.38736*** | 0.1016 |
| Promoters' shares pledged (%) | Management | 0.1587*** | .0582 |
| Earnings Per Share (EPS) | Performance | 0.0275 | 0.0273 |
| Price/Book Value (P/BV) | Performance | -1.8523* | 0.9501 |
| PAT/capital employed | Performance | -0.0541 | 0.1174 |
| Finished goods cycle | Performance | -0.0262 | 0.0325 |
| Return on total assets (ROTA) | Performance | -0.4912*** | 0.1703 |
| Net profit margin (NPM) | Performance | -0.1357 | 0.1656 |
| Size | Control | 1.6650 | 1.2298 |
| Age | Control | -0.4847*** | 0.1053 |
| LR-Statistic (Chi-square) | | 54.58*** | |

Note: This table shows the results of logistic regression in Equation 3. The dependent variable is binary which takes the value of 1 for IBC firms and 0 for solvent firms. The notation '***' refers to significance at the 1% level, '**' refers to significance at the 5% level, and '*' refers to significance at the 10% level.

value of the firm, its cash resources available and the ability of the firm to manage its contingent liabilities. These ratios in the table above are in sync with the management theme.

4. DISCUSSION

Results reveal a significant difference in negative words in MDA between insolvent and solvent firms. The proportion of negative words in compa-

ny MDA reports under IBC is substantially higher than the percentage of words that are unfavorable in the reports of healthy companies. The results thus provide evidence to support the first hypothesis (H_1). It is also clear that the report's tone must be considered when deciding which enterprises are bankrupt and which are not, and it confirms that the negative words from the Loughren and McDonald definition apply in the Indian context. A significant difference in thematic keywords

among the two groups of firms is also observed, which is further validated by a quantitative analysis of panel data. The results strongly suggest using textual analysis to identify insolvency and bankruptcy in firms early. The findings are unique for Indian firms, and applying the mixed method reaffirms the robustness of the results.

The results provide evidence to support the second hypothesis (H_2) of this study, and confirm that certain themes in MDA reports (activity, manageability and performance) can distinguish between IBC and solvent firms before the actual filing of bankruptcy. The negative signs of the coefficients of activity and management scores suggest that fewer keywords in these themes might enhance the likelihood that a business may eventually declare bankruptcy. However, a sudden increase in performance-based keywords (in conjunction with more infrequent activity and management quality-based keywords) may indicate future default. The three significant themes jointly represent over 70 percent of the keywords in MDA reports. Findings further suggest that strong corporate governance as reflected in strong management leads to a higher credit quality and creditworthiness. The results from management quality ratios suggest that higher market capitalization over debt indicates the better financial health of a firm. However, an increase in the debt/equity ratio is a strong indicator of future financial distress. Lowering promoter shareholding is again a strong indicator of future distress and might indicate a loss in promoters' confidence. However, more shares pledged by promoters may indicate an impending crisis in the company.

Highly profitable firms perform better in the market and have low credit risk and high credit quality. The firm's credit rating may be significantly affected by profitability, according to some theories. The coefficient of price to book value, one of the outcomes from the performance ratios, is negative and significant, indicating that a firm's stronger financial situation is indicated by higher equity market performance. The possibility of distress in a company is significantly and negatively affected by return on total

assets, which demonstrates that higher profitability lowers the likelihood of default. The gross fixed asset utilization ratio, among the activity ratios, has a favorable and significant effect on a company's likelihood of defaulting. In the recent past, several steel companies in India expanded their investment in fixed assets in anticipation of higher demand and subsequently increased activity levels. Though, the higher demands never materialized and resulted in unsustainable debt levels² for these companies. The positive sign of the coefficient of gross fixed asset utilization ratio results from several steel companies filing for IBC during the sample period included in the study. It signifies the debt burden and subsequent financial distress these companies faced despite having high asset utilization. The negative sign of the coefficient of export/sales shows a decline in exports as a percentage of sales may indicate a loss in revenues and subsequent financial stress for companies.

The results of the quantitative analysis show that the significant themes in the MDA report can indicate financial distress. The quantitative variables and qualitative analysis of keywords in MDA can be combined to obtain a precise forecast of an impending financial crisis in any company.

From the findings of the study, it can be inferred that comparing the approach taken by regulators to evaluate financial crisis with the tone of financial reports offers advantages. Most financial distress models use some linear combination of financial ratios. Lenders may "window dress" financial data to meet or exceed minimal regulatory standards if they are aware of the financial ratios that regulators use. Additionally, firm managers could propagate false information regarding the caliber of their real estate holdings and regulatory capital through the exercise of their accounting discretion. As a result, in order to forecast financial trouble, regulators or investors may wind up employing erroneous financial data. Extreme events may be difficult to foresee using financial data due to previously unheard-of losses for banks that met or exceeded regulatory standards. (Gandhi et al., 2019).

2 Several news reports highlighted this issue, e.g., <https://www.businessworld.in/article/Steel-The-Big-Fat-Debt-Crisis/20-03-2017-114707/> and <https://www.hindustantimes.com/business-news/npa-crisis-the-rise-and-fall-of-bhushan-steel-in-the-great-indian-debt-trap/story-GHrvRRFIBsMLXKJzbqvaFN.html>

CONCLUSION

The research study explores if the qualitative disclosures as given in the MDA reports of firms in India can preempt firm failure and bankruptcy. For this purpose, word analysis and theme analysis are performed using NVIVO. Words generated from MDA reports are matched with the negative words of the Loughran and McDonald dictionary, and significant themes are identified from the reports. The themes are then validated by applying a mixed-method approach, using quantitative techniques of panel logistic regression.

It is observed that the firms need to go beyond financial information in predicting financial distress and bankruptcy. This study contributes in that it is the first study that has combined quantitative and qualitative approaches in the context of companies under IBC in India. The study contributes by improving upon the existing literature by applying a mixed-method approach in mapping the findings of quantitative research with content analysis. Using two different approaches to confirm findings leads to greater completeness, validity, complementarity (analysis from different perspectives), and generalizability of findings than possible with a single methodology. For managerial implications, this study allows lenders to integrate qualitative variables in their credit risk models.

Textual analysis can help in eliciting additional information over existing financial data thereby enhancing the quantity and quality of financial distress analysis. The textual analysis method could be used by practitioners to complement their analysis of financial data. This could be useful for practitioners in valuing equity and assessing investor sentiment. Auditors can use textual information to identify accounting irregularities and fraudulent activities and predict bankruptcy. To track financial health and identify difficulty, regulators and lenders could examine the frequency of negative phrases and words in all companies' annual reports. In addition to financial variables, greater accuracy in forecasting financial difficulty may be achieved by analyzing the relative frequency of negative terms. As a result, the findings of this study can be used by regulators, auditors, and lenders to create early-warning models that could assist in individually identifying problematic enterprises in the future.

AUTHOR CONTRIBUTIONS

Conceptualization: Vandana Gupta.

Data Curation: Vandana Gupta.

Formal Analysis: Aditya Banerjee.

Investigation: Vandana Gupta, Aditya Banerjee.

Methodology: Vandana Gupta, Aditya Banerjee.

Project administration: Vandana Gupta, Aditya Banerjee.

Resources: Vandana Gupta, Aditya Banerjee.

Software: Aditya Banerjee.

Supervision: Vandana Gupta.

Validation: Aditya Banerjee.

Visualization: Vandana Gupta.

Writing – original draft: Vandana Gupta.

Writing – review & editing: Vandana Gupta, Aditya Banerjee.

ACKNOWLEDGMENTS

The infrastructural support provided by FORE School of Management, New Delhi in completing this paper is gratefully acknowledged.

REFERENCES

1. Abrahamson, E., & Amir, E. (1996). The information content of the president's letter to shareholders. *Journal of Business Finance and Accounting*, 23(8), 1157-1182. <https://doi.org/10.1111/j.1468-5957.1996.tb01163.x>
2. Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589-609. <https://doi.org/10.2307/2978933>
3. Amani, F. A., & Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, 24, 32-58. <https://doi.org/10.1016/j.accinf.2016.12.004>
4. Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405-417. <https://doi.org/10.1016/j.eswa.2017.04.006>
5. Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, 4, 71-111. <https://doi.org/10.2307/2490171>
6. Beaver, W. H. (1968). Alternative accounting measures as predictors of failure. *The Accounting Review*, 43(1), 113-122. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2469909
7. Bonsall, S. B., Leone, A. J., Miller, B. P., & Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3), 329-357. <https://doi.org/10.1016/j.jacceco.2017.03.002>
8. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. Retrieved from <https://psycnet.apa.org/doi/10.1191/1478088706qp063oa>
9. Bryan, S. H. (1997). Incremental Information Content of Required Disclosures Contained in Management Discussion and Analysis. *The Accounting Review*, 72(2), 285-301.
10. Creswell, J. W. (1999). Mixed-method research: Introduction and application. In *Handbook of Educational Policy* (pp. 455-472). Academic Press.
11. Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). An expanded typology for classifying mixed methods research into designs. In A. Tashakkori and C. Teddlie (Eds), *Handbook of mixed methods in social and behavioral research* (pp. 209-240).
12. D'Aveni, R. A., & MacMillan, I. C. (1990). Crisis and the content of managerial communications: A study of the focus of attention of top managers in surviving and failing firms. *Administrative Science Quarterly*, 35(4), 634-657. Retrieved from <https://psycnet.apa.org/doi/10.2307/2393512>
13. Dewasiri, N. J., Weerakoon, Y. K. B., & Azeez, A. A. (2018). Mixed methods in finance research: The rationale and research designs. *International Journal of Qualitative Methods*, 17(1), 1609406918801730. Retrieved from <https://journals.sagepub.com/doi/10.1177/1609406918801730>
14. Donovan, J., Jennings, J., Koharki, K., & Lee, J. (2018). *Determining credit risk using qualitative disclosure*. Available at SSRN 3149945.
15. Engelberg, J. E., Reed, A. V., & Ringgenberg, M. C. (2012). How are shorts informed? Short sellers, news, and information processing. *Journal of Financial Economics*, 105(2), 260-278. <https://doi.org/10.1016/j.jfineco.2012.03.001>
16. Frazier, K. B., Ingram, R. W., & Tennyson, B. M. (1984). A methodology for the analysis of narrative accounting disclosures. *Journal of Accounting Research*, 22(1), 318-331. <https://doi.org/10.2307/2490713>
17. Gandhi, P., Loughran, T., & McDonald, B. (2019). Using annual report sentiment as a proxy for financial distress in US banks. *Journal of Behavioral Finance*, 20(4), 424-436. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2905225
18. Gepp, A., Kumar, K., & Bhattacharya, S. (2010). Business failure prediction using decision trees. *Journal of Forecasting*, 29(6), 536-555. <https://doi.org/10.1002/for.1153>
19. Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(3), 255-274. <https://doi.org/10.2307/1163620>
20. Hanley, K. W., & Hoberg, G. (2010). The information content of IPO prospectuses. *The Review of Financial Studies*, 23(7), 2821-2864. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1327171
21. Haralambie, M. M. (2016). Corporate qualitative and quantitative assessment. *The Audit Financial Journal*, 14(140), 868-868. <http://dx.doi.org/10.20869/AUDITF/2016/140/868>
22. Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication* (1973), 45(4), 363-407. Retrieved from <https://journals.sagepub.com/doi/10.1177/0021943608319388>
23. Hesse-Biber, S. (2010). Qualitative approaches to mixed methods practice. *Qualitative Inquiry*, 16(6), 455-468. Retrieved from <https://journals.sagepub.com/doi/10.1177/1077800410364611>
24. Houghton, K. A. (1988). The measurement of meaning in accounting: a critical analysis of the principal evidence. *Accounting, Organizations and Society*, 13(3), 263-280. [https://doi.org/10.1016/0361-3682\(88\)90004-9](https://doi.org/10.1016/0361-3682(88)90004-9)

25. Hu, N., Liu, L., & Zhu, L. (2018). Credit default swap spreads and annual report readability. *Review of Quantitative Finance and Accounting*, 50(2), 591-621. Retrieved from https://ideas.repec.org/a/kap/rqfnac/v50y2018i2d10.1007_s11156-017-0639-8.html
26. Ingram, R. W., & Frazier, K. B. (1980). Environmental performance and corporate disclosure. *Journal of Accounting Research*, 18(2), 614-622. <https://doi.org/10.2307/2490597>
27. Ingram, R. W., & Frazier, K. B. (1983). Narrative disclosures in annual reports. *Journal of Business Research*, 11(1), 49-60. [https://doi.org/10.1016/0148-2963\(83\)90038-3](https://doi.org/10.1016/0148-2963(83)90038-3)
28. Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 24(4), 602-611. <https://doi.org/10.2307/2392366>
29. Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171-185. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2213801
30. Khemakhem, S., & Boujelbene, Y. (2018). Predicting credit risk on the basis of financial and non-financial variables and data mining. *Review of Accounting and Finance*, 17(3), 316-340. <https://doi.org/10.1108/RAF-07-2017-0143>
31. Kohut, G. F., & Segars, A. H. (1992). The President's Letter to stockholders: An examination of corporate communication strategy. *The Journal of Business Communication* (1973), 29(1), 7-21. Retrieved from <https://journals.sagepub.com/doi/10.1177/002194369202900101>
32. Li, F. (2010). The information content of forward-looking statements in corporate filings – A naïve Bayesian machine learning approach. *Journal of Accounting Research*, 48(5), 1049-1102. <https://doi.org/10.1111/j.1475-679X.2010.00382.x>
33. Liñán, F., & Fayolle, A. (2015). A systematic literature review on entrepreneurial intentions: citation, thematic analyses, and research agenda. *International Entrepreneurship and Management Journal*, 11, 907-933. Retrieved from <https://psycnet.apa.org/doi/10.1007/s11365-015-0356-5>
34. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
35. Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4), 1643-1671. <https://doi.org/10.1111/jofi.12162>
36. Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230. <https://doi.org/10.1111/1475-679X.12123>
37. Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274(2), 743-758. <https://doi.org/10.1016/j.ejor.2018.10.024>
38. Nguyen, B. H., & Huynh, V. N. (2022). Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach. *Journal of the Operational Research Society*, 73(1), 102-121. <https://doi.org/10.1080/01605682.2020.1784049>
39. Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109-131. <https://doi.org/10.2307/2490395>
40. Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning* (No. 47). University of Illinois press.
41. Östlund, U., Kidd, L., Wengström, Y., & Rowa-Dewar, N. (2011). Combining qualitative and quantitative research within mixed method research designs: a methodological review. *International Journal of Nursing Studies*, 48(3), 369-383. <https://doi.org/10.1016/j.ijnurstu.2010.10.005>
42. Previts, G. J., Bricker, R. J., Robinson, T. R., & Young, S. J. (1994). A content analysis of sell-side financial analyst company reports. *Accounting Horizons*, 8(2), 55.
43. Shorten, A., & Smith, J. (2017). Mixed methods research: expanding the evidence base. *Evidence-based Nursing*, 20(3), 74-75. <https://doi.org/10.1136/eb-2017-102699>
44. Shrivastava, S., Jeyanthi, P. M., & Singh, S. (2020). Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. *Cogent Economics & Finance*, 8(1), 1729569. <https://doi.org/10.1080/23322039.2020.1729569>
45. Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), 101-124. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=171436
46. Smith, M., & Taffler, R. (1992). Readability and understandability: Different measures of the textual complexity of accounting narrative. *Accounting, Auditing & Accountability Journal*, 5(4). <https://doi.org/10.1108/09513579210019549>
47. Smith, M., & Taffler, R. J. (2000). The Chairman's statement-A content analysis of discretionary narrative disclosures. *Accounting, Auditing & Accountability Journal*, 13(5), 624-647. <http://dx.doi.org/10.1108/09513570010353738>
48. Swales, Jr, G. S. (1988). Another look at the president's letter to stockholders. *Financial Analysts Journal*, 44(2), 71-73. <https://doi.org/10.2469/faj.v44.n2.71>
49. Tate, W. L., Ellram, L. M., & Kirchoff, J. F. (2010). Corporate social responsibility reports: a thematic analysis related to supply chain management. *Journal of Supply Chain Management*, 46(1), 19-44. <https://doi.org/10.1111/j.1745-493X.2009.03184.x>

50. Tennyson, B. M., Ingram, R. W., & Dugan, M. T. (1990). Assessing the information content of narrative disclosures in explaining bankruptcy. *Journal of Business Finance & Accounting*, 17(3), 391-410. <https://doi.org/10.1111/j.1468-5957.1990.tb01193.x>
51. Thorne, S., Kirkham, S. R., & O'Flynn-Magee, K. (2004). The analytic challenge in interpretive description. *International Journal of Qualitative Methods*, 3(1), 1-11. <http://dx.doi.org/10.1177/160940690400300101>
52. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473-2480. <https://doi.org/10.1016/j.eswa.2007.12.020>
53. Zavgren, C. V. (1985). Assessing the vulnerability to failure of American industrial firms: a logistic analysis. *Journal of Business Finance & Accounting*, 12(1), 19-45. <http://dx.doi.org/10.1111/j.1468-5957.1985.tb00077.x>
54. Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1, 43-52. Retrieved from <https://link.springer.com/article/10.1007/s13042-010-0001-0>
55. Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22, 59-82. <https://doi.org/10.2307/2490859>