Predicting Corporate Distress: A Textual Analysis

The rising corporate debt and higher default rates have led to a continuous increase in distressed loans in Indian financial system. The situation worsened when stressed asset ratio rose from 7.6 % of total advances in March 2012 to 11.5 % in March 2016 and further to 12% in March 2017. Alarmed by the deteriorating asset quality, the Reserve Bank of India (RBI) in April 2015 had urged all commercial banks to put in place an early warning system to prevent financial fraud. The situation had marginally improved during financial years 2018 and 2020. As of March 2020, the total amount of Gross Non-Performing Assets (GNPA) for public sector banks was around Rs. 6.8 trillion (almost $100 billion) down from Rs.8.96 trillion in FY 2018. However, a recent report[1] shows that Bank NPAs may rise to as high as 14.85% of advances by September 2021and this sharp rise in GNPA would be mainly triggered by poor asset quality in public sector banks.

Lenders typically concentrate largely on financial parameters at the time of loan origination and subsequently track the behaviour of borrowers through financial statements and other financial data furnished by the borrower. However, the information in the financial statements may not reveal the actual state of affairs of a borrower.The problem with this approach –generating early warning signals from financial statements- is it may lack predictive power. This would be particularly true for firms which 'window dress' their financial numbers to 'defer' release of bad news.

Much of the research has so far explored the relationship between financial distress and historical accounting information. However, the quantitative financial information comprises only approximately 20% of all the information contained in annual reports (Beattie et al. 2004). Therefore, to obtain a complete picture of financial health of a company, it is necessary that one uses the qualitative information provided in corporate annual reports. There is of late a growing interest among finance and accounting research community in analysing and quantifying the qualitative information present in annual reports. Sunita Goel et al. (2010) study the verbal content and presentation style of the qualitative portion of the annual report using "bag-of-words" approach and suggest that the textual data contains

[1]https://economictimes.indiatimes.com/industry/banking/finance/banking/banks-gross-npa-may-rise-to-13-5-pc-by-sep-2021-rbi-fsr/articleshow/80216967.cms?from=mdr

information that is useful for detecting fraud which is not accurately captured by financial numbers. Loughran and McDonald (2011) analysed the tone of corporate annual reports (sentiment) and observed that sentiments expressed in annual report text data is significantly correlated with profitability, trading volume, and unexpected earnings for listed companies in USA. Fisher et al. (2010) examined the importance of text analytics and information retrieval in accounting, finance and business research. Their findings suggest that developing a computational linguistic tool for accounting and finance research are not straightforward and there is a need for alternative wordlist tailored for finance and accounting domain rather than adapting Harvard Psychological Dictionary developed for psychology and sociology. We have developed a proprietary dictionary for this study.

Realizing the need for greater scrutiny of annual reports, the RBI[2]instructed banks to undertake a detailed study of the Annual Report, and not concentrate merely on financial statements. At present detection of loan frauds takes an unusually long time, which may delay action against any fraudulent entity causing huge losses to financial institutions. So, early detection of any trouble or distress of borrowers would really help in controlling the menace of non-performing assets. The lenders in India should learn the art of extracting information from large text documents and improve their present rating system by supplementing financial parameters with text-based information. This would make the existing rating system more robust.

Our model is developed using text present in the annual report of a company. We have only used three sections of an annual report- Directors Report (including Management Discussion and Analysis), Audit Report and Notes to Accounts. It is important to note that annual report (except the audit report) is a self-report of a company and hence such a document is bound to have strong bias. Yet, we were amazed by the quality of information that one can extract from such a biased text.

[2]Framework for dealing with loan defaults, June 2016

Motivation

For the last 5 years XYZ[3] Limited has been in the top quintile where PD has been increasing over the years, that is the quality of corporate borrowers has worsened over the last five years. Credit rating has remained more or less stable except in FY2020 and performance indicator (Return on Investment) has always been in double digit. So, apparently nothing was alarming for the company.

Table 1: Text-based Probability of Distress and Ratings

Company- XYZ Limited			
FY	PD	Ratings	ROI[4]
2016	0.48	IND BBB+/Positive	17.15%
2017	0.82	IND A-/Stable	16.23%
2018	0.67	IND A/Stable	16.89%
2019	0.81	IND A/Stable	19.74%
2020	0.81	A- and Placed on watch with developing implications (ICRA) (September 2019)/ IND A Rating Watch Negative (October 2019)	14.90%

However, the annual report of the company over the past five years have highlighted several concerns and indicated signs of distress. The material information captured in the text of the annual report, in this example, proves that it makes economic sense to analyse the non-financial information as seriously as one does for financial information. We find that directors' report provide most of material information and audit report provided least marginal information. It appears that stock market has priced the credit risk much before the rating agencies did. The share price of the company, which was Rs. 76 in March 2016,

[3] Actual Name of the company withheld.

[4] ROI or Return on Investment is calculated as a percentage of Operating Profit divided by Invested capital, where Invested capital is the sum of Net worth and Debt of the company. Data source: Prowess

reached a peak in January 2018 at Rs. 137 a piece. It then nosedived during 2019 and early 2020 to reach at an all-time low of Rs. 17 in March 2020. The PD scores (Table 1) also captured similar trends- the score did improve in 2018 and then collapsed in next two years. Some examples of disclosures in the Annual Report of the company are provided below:

- The subsidiary XYZ infrastructure holdings limited which has invested in equity shares and / or in preference shares or has advanced monies in some companies have incurred losses during the year and also have accumulated losses as at the end of the reporting period. (*FY 2016*)
- The demonetization announced by the union government in November, 2016 and the consequent slowdown in the economy in the second half of the financial year resulted in decreases in the turnover posted and the net profit earned by the company as compared with the previous year. (*FY 2017*)
- The subsidiary had accumulated losses (excluding foreign currency transition reserve) of 879.46 million. (*FY 2018*)
- Cancellation of Rs. 6,100 crore worth orders (*FY 2020*)

This example motivates us to develop a text-based analytical tool that would predict corporate distress. The PD tool of the Textplor is an outcome of our efforts.

Data and Methodology

Our initial data-set consisted of annual reports of both public and privately held companies operating and registered in India. We have selected the companies functioning in around 36 different sectors except financial and insurance sector. Due to special business nature and financial structure, insurance and banking sector firms are excluded.

We followed two-stage approach to estimate a probability of distress (PD) from text of any annual report. First, we used natural language processing (NLP) to extract sentiment scores. Second, we used a logistic regression to convert sentiment scores to a probability estimate of distress, which we call PD.

Texts are extracted from the annual reports and sentences are separated using separators for sentence-level sentiment analysis. Sentences are used instead of words as the text can contain a number of positive as well as negative words, which are incapable of providing the true sentiment of the text. Next, NLP (Natural Language Processing) is used for data preparation that does not understand text. So, sentences are transformed into column vectors with each word given a numerical value using a trained model. But, the length of the sentence vector is

not equal as the sentence can be of 10 words or 50 words long, for which an algorithm is used, that merges words and includes some extra steps. Equal sentence vectors are then concatenated with other sentences to form a matrix with thousands of sentences. Some binary features are added to this matrix, related to sentiments, for better analysis using Support Vector Machine (SVM) Model. This data is divided into two parts: training set from the year 2009-2013 with 90% of the data and test set for the years 2013 to 2015 with 10% of the data. After training, the test result categorizes each row into three buckets: *Positive, Neutral* and *Negative*, by giving a number from -1, 0 and 1. The scores are used as sunshine (positive) and fear(negative) scores with an accuracy of 80.7% for a dataset of 700 companies. Distress Intensity (DI) is calculated using the ratio of Fear by Sunshine scores. If this value is greater than 1, it implies fear is more than sunshine and is tending towards negative sentiment and vice versa. Defaulted companies were given a value of 1 and others 0, for each year from 2010 to 2020. Using this as the training model, DI is used as the test model for conducting Logistic Regression. Coefficients from the Logistic Regression provide the Probability of Default (PD) which can be classified into 4 categories: High PD but not defaulted, High PD and defaulted, Low PD and not defaulted and Low PD but defaulted. This classification provides some threshold values – 0.3 and 0.5, that classifies PD into 3 categories: low, moderate and high PD with an accuracy above 75%.

We have looked at the effect of PD on the credit rating and operating performance of firms. Since, PD is available on an annual basis; the above variables are estimated annually. Credit Rating had various categories from 'highest safety' to 'default'. We have categorised them numerically on a six-point scale, ranging from 0 to 5 with categories 'default', 'high risk', 'inadequate/substantial risk', 'moderate/adequate safety', 'high safety' and 'highest safety' respectively. To measure financial performance, Return on Investment (Operating profit by total Invested Capital) and Debt-Equity ratio (Debt by Equity) are considered.

We have considered annual reports of 292 companies for which we have continuous data for the entire sample period (from financial year 2010 to 2020).

Table 2: Descriptive Statistics

MEAN										
Deciles	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
	Panel A: PROBABILITY OF DISTRESS									

1st	0.456	0.554	0.542	0.545	0.685	0.617	0.815	0.823	0.827	0.835
2nd	0.270	0.303	0.306	0.325	0.413	0.396	0.498	0.485	0.568	0.566
3rd	0.234	0.248	0.249	0.267	0.304	0.305	0.360	0.364	0.403	0.433
4th	0.207	0.208	0.219	0.230	0.259	0.248	0.302	0.300	0.328	0.356
5th	0.186	0.188	0.196	0.210	0.226	0.215	0.264	0.257	0.281	0.306
6th	0.170	0.174	0.176	0.192	0.202	0.190	0.230	0.226	0.242	0.264
7th	0.158	0.159	0.156	0.173	0.185	0.168	0.206	0.197	0.217	0.217
8th	0.142	0.141	0.140	0.155	0.159	0.152	0.174	0.166	0.191	0.161
9th	0.089	0.041	0.070	0.133	0.081	0.115	0.134	0.122	0.151	0.004
10th	0.000	0.000	0.000	0.032	0.000	0.000	0.000	0.000	0.032	

Panel B: CREDIT RATING

1st	3.391	3.043	3.043	2.692	2.435	2.435	2.696	2.304	2.409	1.500
2nd	3.875	3.542	3.542	3.720	3.261	3.120	2.923	3.375	3.542	3.074
3rd	3.429	3.654	3.654	3.519	4.960	3.560	3.500	3.840	3.357	3.846
4th	3.560	3.773	3.773	3.462	4.154	4.120	3.724	3.556	3.786	4.111
5th	4.077	3.786	3.786	4.083	3.769	3.962	4.185	3.643	3.931	3.862
6th	4.036	4.120	4.120	3.889	3.704	3.741	3.893	4.037	4.000	3.885
7th	3.750	3.893	3.893	4.087	3.607	4.143	4.143	5.286	4.214	4.143
8th	4.200	4.040	4.040	4.077	4.111	4.107	3.964	4.080	3.923	3.862
9th	3.875	3.692	3.692	4.115	4.000	3.640	3.926	3.857	3.963	4.107
10th	3.222	3.333	3.333	3.360	3.538	3.240	3.296	3.241	3.320	3.333

Panel C: DEBT-EQUITY RATIO

1st	-0.218	2.067	2.708	8.376	9.655	2.466	3.791	3.084	1.582	0.829
2nd	1.603	1.416	0.413	2.666	2.434	2.338	0.351	2.169	1.579	1.843
3rd	2.095	0.414	1.143	1.299	1.639	1.389	0.673	0.800	1.493	0.929
4th	1.071	2.402	1.001	1.142	1.011	0.947	1.046	0.900	1.349	0.568
5th	1.153	0.937	0.999	1.294	0.754	0.857	0.820	1.455	1.370	-2.945
6th	0.743	0.836	1.540	1.156	1.101	1.139	0.919	0.313	0.532	0.362
7th	1.207	0.605	1.196	0.732	0.832	0.626	0.882	0.563	0.549	0.917
8th	0.520	0.984	0.877	0.972	0.549	0.850	0.867	1.373	0.642	0.346
9th	0.766	1.564	0.712	0.904	1.389	0.567	1.017	1.253	0.809	-7.921
10th	1.031	2.372	0.711	1.246	1.371	-0.089	24.693	1.957	1.102	

Panel D: Return of Investment										
1st	13.930	8.496	8.452	12.212	10.512	4.856	-27.843	-14.093	-11.666	-50.814
2nd	16.648	13.334	11.641	12.109	13.696	12.377	11.122	12.798	7.186	6.208
3rd	13.100	18.900	15.608	17.529	17.974	18.219	17.065	15.030	111.845	16.265
4th	11.943	16.735	12.394	18.270	17.206	16.577	15.270	16.041	19.468	17.889
5th	16.569	17.411	16.362	19.211	12.823	17.830	25.201	14.912	21.130	19.084
6th	20.298	15.848	18.639	14.816	17.843	15.257	17.821	20.575	17.926	13.350
7th	21.376	18.281	15.769	13.881	16.347	15.678	15.737	15.988	18.960	20.762
8th	22.349	25.929	18.863	20.636	15.964	15.517	15.411	13.055	12.228	23.365
9th	24.265	17.887	18.674	17.335	19.306	20.909	16.059	17.208	16.814	12.863
10th	11.757	15.124	15.040	14.128	15.929	14.639	16.339	12.365	13.599	8.837

Table 2 shows mean values of the dependent and independent variables of interest. The top(bottom) decile represents 10 percent of the companies which have highest(lowest) PD. Results show that quality of corporate borrowers has worsened over the last ten years with the average PD for the worst of the companies has almost doubled during this period. The pattern is similar for every decile. The companies in the deciles for other panels (Table 2) are same as in Panel A. In other words, the mean credit rating score for companies with highest PD (decile 1) for the financial year 2020 was 2.5 (between substantial and high-risk categories), and the mean ROI for companies with highest PD (decile 1) for the financial year 2020 was -50.81%. However, the pattern is not that striking between PD scores and Debt-equity ratios.

Since our data consist of cross-section variables changing across time; panel regression is used. With weakly balanced data (each panel contains the same number of observations but not the same time points); a panel is set, using the 'Uniquid' (that takes value from 1 to 292 for 292 companies) as the cross-section variable and 'Year' (from 2010 to 2020) as the time variable.

To calculate the effect of PD on Credit rating and ROI, we run three separate sets of panel regressions.

A. Regressions with ROI(t), ROI(t+1) and ROI(t+2) as the dependent variables.

$ROI_{it} = \alpha + \beta_1 PD_{it} + \beta_2 ROI_{i,t-1} + \varepsilon_{it}$.....[eq A1]

$ROI_{i,t+1} = \alpha + \beta_1 PD_{it} + \beta_2 ROI_{i,t} + \varepsilon_{it}$.....[eq A2]

$ROI_{i,t+2} = \alpha + \beta_1 PD_{it} + \beta_2 ROI_{i,t+1} + \varepsilon_{it}$.....[eq A3]

B. Regressions with Debt-Equity(t), Debt-Equity(t+1) and Debt-Equity(t+2) as the dependent variables.

$Debt - Equity_{it} = \alpha + \beta_1 PD_{it} + \beta_2 Debt - Equity_{i,t-1} + \varepsilon_{it}$.....[eq B1]

$Debt - Equity_{i,t+1} = \alpha + \beta_1 PD_{it} + \beta_2 Debt - Equity_{i,t} + \varepsilon_{it}$.....[eq B2]

$Debt - Equity_{i,t+2} = \alpha + \beta_1 PD_{it} + \beta_2 Debt - Equity_{i,t+1} + \varepsilon_{it}$.....[eq B3]

C. Regressions with Credit Rating(t), Credit Rating(t+1) and Credit Rating(t+2) as the dependent variables.

$CreditRating_{it} = \alpha + \beta_1 PD_{it} + \beta_2 CreditRating_{i,t-1} + \varepsilon_{it}$.....[eq C1]

$CreditRating_{i,t+1} = \alpha + \beta_1 PD_{it} + \beta_2 CreditRating_{i,t} + \varepsilon_{it}$.....[eq C2]

$CreditRating_{i,t+2} = \alpha + \beta_1 PD_{it} + \beta_2 CreditRating_{i,t+1} + \varepsilon_{it}$.....[eqC3]

Results and Analysis

We observe (Table 3) that PD is negatively correlated with Credit Ratings and ROI. High PD means higher probability of distress, which obviously means credit ratings would fall and so will the return on investment (ROI). Also, PD is positively correlated with Debt-Equity ratio, which is again ideally correct.

Table 3: Correlation between PD, Credit rating(t), Credit rating(t+1), Credit rating(t+2), ROI(t),ROI(t+1), ROI(t+2), Debt-Equity(t), Debt-Equity(t+1) and Debt-Equity(t+2).

Correlation	PD	Credit rating (t)	Credit rating (t+1)	Credit rating (t+2)	ROI (t)	ROI (t+1)	ROI (t+2)	Debt-Equity (t)	Debt-Equity (t+1)	Debt-Equity (t+2)
PD	1.000									
Credit rating(t)	-0.155	1.000								

Credit rating(t+1)	-0.159	0.593	1.000							
Credit rating(t+2)	-0.146	0.556	0.615	1.000						
ROI(t)	-0.130	0.271	0.280	0.288	1.000					
ROI(t+1)	0.004	0.053	0.038	0.008	0.155	1.000				
ROI(t+2)	-0.027	0.085	0.092	0.080	0.152	-0.009	1.000			
Debt-Equity(t)	0.032	-0.096	-0.099	-0.093	-0.051	0.023	-0.127	1.000		
Debt-Equity(t+1)	0.107	-0.059	-0.099	-0.099	-0.039	-0.016	0.014	0.096	1.000	
Debt-Equity(t+2)	0.035	-0.046	-0.032	-0.063	-0.019	-0.009	-0.008	0.019	0.089	1.000

There are two types of effects in a panel model: Fixed or Random effect. A model is a fixed effect model if the variables are constant across individuals, random otherwise. To check which effect is best suited for the given data, we have conducted the Hausman test. Results show that it is better to use random effect models.

Also, to correct for Autocorrelation or Heteroscedasticity, if any, the Feasible Generalized Least Square (FGLS) Panel model is used. But since the coefficients of FGLS and normal panel model are the same, we can conclude that there is no problem of Autocorrelation or Heteroscedasticity and the normal random effect model is used. All the panel regression results are shown in Table 4.

Table 4: Random Effect Panel Regression on PD(t) with Credit Rating (t), Credit Rating (t+1), Credit Rating (t+2), ROI(t), ROI(t+1) and ROI(t+2), Debt-Equity (t), Debt-Equity (t+1) and Debt-Equity (t+2)as the dependent variable

	Dependent Variable

	PANEL A			PANEL B			PANEL C		
	i.ROI(t)	ii.ROI(t+1)	iii.ROI(t+2)	i. Debt-Equity (t)	ii. Debt-Equity (t+1)	iii. Debt-Equity (t+2)	i. Credit rating(t)	ii. Credit rating(t+1)	iii. Credit rating(t+2)
Constant	22.36***	20.60***	19.52***	.58	-.59	.43	2.04***	1.89***	1.81***
PDt	-29.73***	-25.19***	-22.50***	2.85*	7.91***	3.89**	-1.06***	-.84***	-.71***
ROI(t-1)	.031								
ROI(t)		.028							
ROI(t+1)			.029						
Debt-Equity (t-1)				.092***					
Debt-Equity (t)					.091***				
Debt-Equity (t+1)						.091***			
Credit rating(t-1)							.52***		
Credit rating(t)								.54***	
Credit rating(t+1)									.55***

**** represents significance at 1%, ** at 5% and * represents significance at 10%*

Results in Table 4 clearly show that PD significantly explains operating parameters, degree of leverage and credit rating of firms. The coefficients of PD are negative and significant for credit rating and ROI. The lagged terms of dependent variables are used in the regression to take care of autocorrelations in the variable caused by stickiness. Even after controlling for lag terms, PD significantly explains variations in the dependent variables.

Theoretically, Return on Investment (ROI) decreases as distress intensity increases. PD has predictive powers. For example, current-year PD explains ROI two years hence. In other words, not only the PD of any year could explain operating performance of a firm two years later, the negative coefficient remains equally significant.

The coefficient of PD is positive and significant in explaining leverage of any firm. Higher the PD, higher should be the leverage (Debt-equity ratio). One may argue that the leverage should affect the PD and not the other way round. However, we have observed in our sample that PD precedes leverage. Firms having better profitability have lower PD and lower leverage. Whereas firms which start facing operational stress would tend to borrow more to fund growth or operations. PD captures the stress much before firms start increasing leverage. That is why, we observe (Table 4 Panel B)PD of any given year explains debt-equity two years forward. Either debt increases or equity decreases as PD increases.
Similar explanations can be put forward for credit rating as dependent variable. Our anecdotal evidence (table 1) has shown that PDs precede ratings. Results in Table 4 (Panel C) further corroborates the anecdotal evidence. PD is negative correlated with the ratings. Secondly, the coefficients of PD are negative and significant for ratings in subsequent two years. Therefore, companies reveal value-sensitive information in the Annual Report and it takes some time for the rating agencies to incorporate such information in their ratings.

Therefore, it seems that PD tool is useful for the following reasons: (a) it automatically generates distress probability using textual information of any annual report; (b) it has a predictive power for eventual downgrade/upgrade in credit ratings; (c) corporate financial statements reflect financial stress much later whereas PD captures it in advance. Hence, corporate lenders, rating agencies and asset managers should use the PD tool to augment their respective models to ensure better decision making capabilities.

References

Beattie, V., McInnes, W., &Fearnley, S. (2004). A methodology for analysing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes. Accounting Forum, 28 (3), 205–236.

Fisher, Ingrid, E., Margaret, R. Garnsey, Sunita Goel, and Kinsun Tam . (2010). The Role of Text Analytics and Information Retrieval in the Accounting Domain, Journal of Emerging Technologies in Accounting , 7(1), 1–24.

Goel Sunita, Jagdish Gangolli, Sue, R. Faerman, and OzlemUsuner. (2010). Can Linguistic Predictors Detect Fraudulent Financial Filings? Journal of Emerging Technologies in Accounting, 7(1), 25-46.

Tim Loughran and Bill McDonald(2011), When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. Journal of Finance, 66 (1), 35-65