NIBM WORKING PAPER SERIES

Loan Frauds and Bad Boy Billionaires: A New Approach of Loan Fraud Prevention using Natural Language Processing (NLP)

Smita Roy Trivedi
Dipali Krishnakumar
Richa Verma Bajaj

Working Paper
(WP04/2021)

NATIONAL INSTITUTE OF BANK MANAGEMENT
Pune, Maharashtra, 411048
INDIA
November 2021

Citation Guideline:
Roy Trivedi, Smita, Dipali Krishnakumar and Richa verma Bajaj (2021), "Loan Frauds and Bad Boy Billionaires: A New Approach of Loan Fraud Prevention using Natural Language Processing (NLP)". NIBM Working Paper Series, WP04/November.
https://www.nibmindia.org/static/working_paper/NIBM_WP04_SRTDKRVB.pdf

Loan Frauds and Bad Boy Billionaires: A New Approach of Loan Fraud Prevention using Natural Language Processing (NLP)

Smita Roy Trivedi, Dipali Krishnakumar and Richa verma Bajaj

NIBM Working Paper No. 04
November 2021

ABSTRACT

Credit frauds, where a loan turns into bad debt on account of fraudulent activities damage the core business of the banking industry and dent its reputation. While transactionsbased activities like credit card frauds or cybercrimes are captured at a point in time, credit frauds perpetuate over time. We present a methodology to analyse frauds happening over time through NLP tools. The methodology can use linguistic information on customers as well as documents to identify causes of frauds. We use the methodology to analyse 653 known cases on fraud from India, which has seen a growing number of credit frauds, to identify the Early Warning Signals. We develop a ranking of the EWS and further use an ordered logit model to analyse the most important EWS impacting high value frauds.

Keywords: Credit frauds, Scoring of Early Warning Signals, Survey, Optical Character Recognition, Natural Language processing (NLP), Logistic Regression, Operational Risk

JEL Classification Number: C83, C88, G21, M48

Acknowledgements: The project was undertaken as a part of the 'RBI Scholarship Scheme for Faculty Members of Academic Institutions 2020'The authors place on record their gratitude to the bankers from commercial banks in India who participated in the survey and whose challenges in EWS identification motivated the study.
The research assistance provided by Shivam Kumar and Tanvi Paleja, students of NIBM PGDM (B&FS) was exceptional and deeply acknowledged. We also thank Shivam for his feedback on NLP program code which helped to refine it.
We gratefully acknowledge the support from Krishnakumar Ramanujam, EVP and Country Head, Abzooba Infotech India Private Limited for the OCR code as also the feedback and crucial modifications to the NLP code, which reduced the run time hugely. We thank Aditya Aggarwal, Practice Lead Advanced Analytics, Abzooba Infotech India Private Limited for an initial feedback on the code.
We thank Mukund Ladha, Rhythm Kumar, Riya Nadkarni, Sunil Kumar Vashisth, Vikrant Sharma, students of NIBM PGDM (B&FS) for the validation/ human audit.
On behalf of all authors, the corresponding author states that there is no conflict of interest.

Smita Roy Trivedi ***(Corresponding Author)***
National Institute of Bank Management, India
E-mail: smita@nibmindia.org

Dipali Krishna Kumar
National Institute of Bank Management, India
E-mail: k.dipali@nibmindia.org

Richa Verma Bajajn
National Institute of Bank Management, India
E-mail: rica@nibmindia.org

Loan frauds and Bad Boy Billionaires: A new approach of loan fraud prevention using Natural Language Processing (NLP)

1. Introduction

"Mounting evidence of massive frauds involving the loss of billions of dollars provoked an angry public to demand answers to tough questions: Who stole all the money? Why aren't they in prison? How much of the money can we get back?"

(Calavita, et al., 1997 on United States Saving and Loan crises of 1980s)

"Who killed our banks? Rs 9.5 lakh crore is a lot of NPAs. And India's public sector banks account for nearly 80% of it. How did it come to this? And what kind of systemic fixes will it take to break the endless cycle of bad debts and bailouts?",

(India Today on the Indian banking frauds of 2000s, May 18, 2018)

Loan scandals and credit fraud weigh heavily on the banking system. Unlike corporate frauds which are majorly on 'behalf of the corporation' (Calavita et al., 1997), credit or loan frauds hint an insider collusion that destroys the banking system, a greater shame. India's loan fraud story is a case in point. The loan frauds have increased by 28 per cent by volume and 159 per cent by value during 2019-20[1]. Loan frauds, unlike other transaction-based frauds are not one time and usually take years to show up as a 'fraud'. What starts out as normal lending turns into a bad loan with the money siphoned off over a period of time. The transaction-based alerts alone therefore do not work in these cases. As these frauds typically have a gestation time and throw up the early warning signals or red flags over time, banks are given a list of indicators which can detect a brewing fraud. These guidelines on early detection of fraud, known variously as 'Early Warning Signals' (EWS) or 'Red Flag Indicators'(RFI). In India, the central bank, Reserve Bank of India has issued a comprehensive set of Early Warning Signals (RBI Master Directions on Frauds – Classification and Reporting, 2017) [2].

While weak implementation and non-identification of Early Warning Signals (EWS) on time is an important reason behind delay in detection of frauds by lenders (RBI Annual Report, 2020), literature suggests *all early warning signals* may not be equally important (Smith et al. 2005). Moreover, bankers have to tread a fine line between the EWS identification and normal business activities to deal with rising fraud losses. In such cases, how do bankers make a judgement? To address this lacuna, we present a methodology to analyse known frauds using NLP tools and identify the most relevant EWS. We develop a ranking of the EWS and further use an ordered logit model to analyse the most important EWS impacting specifically high value frauds. The analysis of EWS

[1]98 percent of the frauds are in credit, resulting in high NPAs in banks (Charan et al. 2016). In the similar direction, Bajaj and Krishnakumar (2020) found that diversion of funds and non-identification of early warning signals are the major causes of rising NPAs in banks in India.

[2] RBI issued the Master Direction on Frauds[2] on July 01, 2016, revised and updated in July 2017, for commercial banks and select Financial Institutions in India, which provides an illustrative list of 43 Early Warning Signals (EWS) for alerting bank officials about the wrongdoings in the loan accounts before they turn fraudulent

categorised is mapped to a finer and broader category, Category 1 and 2 respectively as explained in Section 3 and detailed in Appendix 1.

The use of Red Flags for fraud detection is well accepted in literature (Albrecht et al., 2012; Coenen 2008; Stamler et al., 2014) and a host of studies have given a ranking of EWS (See Smith et al. 2005 for a survey). However, the method used for such scoring is generally survey based as in Smith at al. (2005). Survey output is influenced by expert's judgement, which may not always be unbiased (See for example, Koehler et al. 2002). Our study uses a survey for preliminary understanding and scoring of EWS covering bankers from credit, international banking, risk management or audit departments. Our results show that while 96% of the EWS suggested by RBI are effective, bankers find that only 36% of EWS are easy to identify (complete results in Appendix 3).

We therefore focus on the data available on what actually happened in cases of fraud. While India has a growing number of fraud cases, there is no publicly available data on frauds[3]. This data constraint drives us to work on a unique methodology to tap a hitherto unused information source: cases registered with Central Bureau of Investigation (CBI)[4]. While fraud detection techniques require relevant, and clean data, such data is scarce (Phua et al.,2010), majorly because organisations are hesitant to disclose fraud data on account of legal and competitive pressures. Use of NLP technique, as done in this paper, can be a useful methodology is this case.

We find that in Category 1, Diversion of Funds, Concealment or Falsification of documents and Internal Control Weaknesses are the most important EWS for the cases analysed. For Category 2 we present a ranking of all EWS based on normalized scores. This ranking is frequency based without a consideration of fraud value. However, we contend that the same set of EWS may not be relevant for all frauds, ranked by their magnitude. We therefore use an ordered logistic model to analyse linkages between the fraud value and EWS for cases where fraud values are available. We find Inter-Group/Concentration of Transactions, Issues in Primary/Collateral Security, Operations in Accounts, Diversion of Funds, Internal Control Weaknesses, Other Signals, and Regulatory Concerns have statistically significant impact on fraud quantum category.

We contribute to the existing literature in three important ways. First, to the best of our knowledge, this is the first study which develops a ranking or scoring of either EWS/RFI on the basis for NLP tools. Given the phenomenal increase in the losses to banks on account of loan frauds, this will give bankers a useful operational tool to understand which EWS occur frequently and which are more important for higher fraud values. It will provide the regulators and banks an understanding of where loopholes in the monitoring could be and how to address them by making use of transaction based/account wise/bank specific data. Secondly, we use a unique methodology for identification of EWS based on NLP techniques, which makes possible the harnessing of a rich source of data, not so far attempted, to the best of our knowledge. The technique can use information over a period of time to pre-empt cases of fraud. Third, NLP methodology creates a rich Dictionary containing words and phrases that can be used operationally to identify EWS.

[3] RBI set up the Central Fraud Registry (CFR) on July 3, 2017 to enable 'early detection mechanisms' for frauds as well as 'a searchable centralised database for use by banks' (RBI, 2016). However, the data is not publicly available.

[4] CBI is the Government of India's nodal investigative agency with a vision of "Combating corruption in public life, curb economic and violent crimes through meticulous investigation and prosecution".

The rest of paper is organized as follows. Section II reviews the literature. Section III delves into the Data and Methodology, Section IV details the results and Section V discusses and concludes.

2. Review of Literature

It is accepted that availability of red flag indicators/ early warning signals can help law enforcement agencies to identify fraud timely and prevent future losses. Availability of red flag indicators make enforcement agencies more aware of its existence. The use of Red Flags for fraud detection is common (Albrecht et al., 2012; Coenen 2008; Stamler et al., 2014) and more the 'red flags' in a business, the more likely is a fraud taking place (Burns 1997).

Are the RFIs/EWS effective? Pincus (1989) finds no significant difference in the assessed risk of fraud by questionnaire users and non-users for a no-fraud case, so that questionnaire use had no significant impact on fraud risk assessment. More worryingly, use of questionnaire was dysfunctional for the fraud case and the non-users outperformed the questionnaire users in these cases. This suggests that merely having a RFI/EWS list may not be enough. In fact, understanding which of the RFI/EWS are more important is essential.

Bell and Carcello (2000) provide a comprehensive set of significant, and insignificant, risk factors. Smith et al (2005) point out that even though long lists of red flags have been provided in the US Statement on Auditing Standards (SAS), but they do not have a guide to weight or prioritise fraud risk factors. Assessment of fraud risk will be more effective if the EWS can be distinguished based on risk factors and weights (Shelton et al., 2001). Further, a consensus on red flag indicators is essential. Heiman-Hoffman et al., (1996) point out that auditors' consensus on RFIs makes them easier to spot.

Weisenborn and Norris (1997) applied 86 red flags to 30 fraud cases and found important red flags were "dishonest or unethical management" and "inadequate internal controls or failure to enforce controls". Loebbecke et al. (1989) used RFIs from SAS No. 53 and found that red flags included in the survey had high frequency of occurrence. Apostolou et al. (2001) found "known history of securities law violation", "significant compensation tied to aggressive accounting practices" and "management's failure to display appropriate attitude about internal control" as the three most important red flags rated. Heiman-Hoffman et al. (1996) found that auditors perceived "attitude factors" such as dishonest, hostile, aggressive and unreasonable management attitudes to be more important warning signs than "situational factors". In contrast, Abdul Majid and Tsui (2001) revealed situational factors such as "difficult to audit transaction" and "indication of going concern" to be more important than "attitude factors".

Bell and Carcello (2000) find 'management lied to the auditor'; 'a weak internal control environment'; 'an unduly aggressive management attitude'; 'undue management emphasis on meeting earning projections'; and 'significant difficult-to-audit transactions' as important. Beasley et al. (2000) compare known fraud and no-fraud benchmarks and find companies who exhibited fraud had fewer audit committees, fewer independent audit committees, fewer audit committee meetings, less frequent internal audit support and fewer independent board members. Smith et al. (2005) find "management failure to display appropriate attitude toward internal control" to be the most important individual red flag. Ghafoor et al. (2019) identifies the key factors that led to financial reporting

fraud among companies in Malaysia using data on a sample of 76 firms that had committed financial fraud for the period of 1996–2016.

Fraud analytics can benefit hugely from machine learning models (Jullum et al. 2020, Singh & Best 2019, Kannan and Somasundaram 2017). However, these studies are suited to analysing transaction at a particular point of time and do not cover the credit frauds that happen over time.

To address these limitations, we tap into the only source of information on fraud publicly available and use a unique methodology for accessing the information, as detailed in the next section. We also present in the results section a comparison of finding from survey and NLP technique. It is encouraging to note survey results at Category 1 level are broadly similar to our findings, but there is a lack of granularity and a mismatch at category 2 levels, which suggest our methodology of case-based analysis is useful.

3. Data, Methodology and Sources

1. Survey:

The objective of this study is to develop a scoring model to identify frauds in Banks in India. The Earning Warning Signals (EWS), given by RBI through the Master Circular on Frauds, July, 2017 form the basis of questionnaire design. The objective of the questionnaire was to analyse two important aspects: (i) effectiveness of RBI EWS and (ii) ease of their identification, through survey-based approach. We get 63 responses on first questionnaire and 48 on the second questionnaire from bankers working in credit, forex, risk and internal audit department of the bank. In order to gauge the significance attached by the respondents to Early Warning Signals, the responses were obtained on five-point scale ranging from 1 to 5. Here, 1 signifies "very rarely effective/very difficult to identify" and five signifies the "very effective/very easy to identify" for questionnaire I and Questionnaire II, respectively. In order to assess the reliability of the scale of the survey questions, the Cronbach alpha was computed. Nunnally (1978) recommends a minimum level of 0.7 for Cronbach's alpha. We found Cronbach alpha value of 0.95 and 0.96 for Question I and Questions II, respectively, which is satisfactory. The results of the analysis are presented in Section IV (1).

2. Natural Language Processing (NLP) and quantification of the Early Warning Signals (EWS)

i. Data and sources

Publicly available CBI files from various departments are used, which record the details of the case which can help to identify the EWSs that happened in each case. The final numbers of credit fraud cases extracted from the different branches of CBI are listed in Table 1

The CBI FIRs include chronological description of fraud event and on how the fraud was committed (tapped for ranking of EWS) and additional details about the bank reporting the incident, amount of fraud, details of the borrower, start and end date of fraud, business of the company, nature of the organisation, type of facility used, banking arrangement. On each of the parameters listed above we have information on number of fraud incidences (frequency) and magnitude of loss (severity). The parameter wise analysis is presented in Section IV (2).

These FIRs are an authentic data source to carry out our analysis, albeit with one challenge. Though the documents are digital documents in PDF form, they are typically scanned copies of letters, reports etc. Considering that the number of reports required for a ranking, manual reading and analysis of each report is not practical.

We have used an Optical Character Recognition (OCR) code which reads each file and converts to machine readable text. While there are few sections of the FIRs which may not be converted into text, for example tables, hand written script, script in regional languages etc, we find that the major portions of the FIR (which includes a letter from the bank with a description of the incident) are converted to a machine-readable text format making it amenable to further analysis using Natural Language Processing. The code is provided in Appendix 2 a.

ii. Training for Analysis and construction of Dictionary

Dictionaries used to match the phrases with the corresponding EWS is the crucial ingredient for any NLP study. NLP studies using customised dictionaries have been frequently used in equity market sentiment scoring, understanding of company financials and central bank communication. The success of Dictionary and frequency-based analysis depends to a large extent on the 'Dictionary' or collection of words being searched for. Using pre-built dictionaries: pre—defined positive and negative word categories as in the Harvard-IV-4 psychosocial dictionary is common (Tetlock et al. 2008). However, the problem with such pre-built dictionaries is that words are used contextually and therefore same set of words classified in one context may not have the same interpretation in another. For example, the negative words used in Harvard-IV-4 psychosocial dictionary may not be suitable or represent negative connotation in a financial context (Loughran and McDonald, 2011). This underlines the need for a customised dictionary to be used. As there is no pre-built dictionary for fraud analysis, we work on building a customised one. The customised dictionary is then use for analysing of the test content.

For creating the customised dictionary, we use 50 files for the training. Each of these cases are read by one of the authors and first set of EWS pertaining to each is identified. In the next round, all EWS for each case are discussed by the authors, and vetted or weeded out based on the discussion. The authors bring together their expertise in operational risk, corporate credit, international banking and fraud for identification of the set of EWS for each case. Table 2 gives a list of words and EWS it is mapped to as an example. The first round of analysis by authors on 50 cases yielded a list of around 1500 words, of which 1000 were retained post the first round of vetting by authors. We use this customised dictionary to pick up phrases and categorize them into defined EWS buckets as detailed in the next section.

iii. NLP Methodology

We follow a NLP based text classification methodology to match words appearing in the cases registered with four law enforcement cells of the Central Bureau of Investigation (CBI Branches, Banking & Securities Fraud Cell, Economic Offence Wing and Anti-Corruption Bureau) and mapped them with the list of EWS based on the Dictionary created. NLP focuses on the development of appropriate tools to make computer systems understand human languages and do desired analytics. Applications of NLP include

language text processing, classification and summarization which is used in this study in order to categorise different words and phrases used in the FIR reports into EWS buckets for each case. The program code is provided in Appendix 2 b.

iv. EWS mapped to two categories

The analysis of EWS categorised is mapped to a finer and broader category, Category 1 and 2 respectively. The Institute of Chartered Accountant of India (ICAI) have classified the RBI's Early Warning Signals under 7 broad categories, (i) Operations in Account; (ii) Concealment or Falsification of Documents; (iii) Diversion of Funds; (iv) Issues in Primary/Collateral Securities; (v) Inter-Group/Concentration of Transactions; (vi) Regulatory Concerns and (vii) Other signals. We denote this broad classification as Category 1. The Category 1 classification provides an understanding of the risk profile of Bank for fraud detection and control. It is not only useful for regulatory reporting, but also provides an important picture for top-management oversight.

The classification of EWS under Category II provides a granular picture of these signals, which provides a greater understanding and ease of identification to the branch officials/credit officers/ forex officers/internal auditor, to relate and report such incidences to the reporting authority of their respective bank. While analysing the cases for training the Dictionary we came across other EWS which were not covered in RBI list but were important in the particular case. These were added and, in some cases, a new classification under EWS Category 1 was created for the same. The list of Category 2 EWS, as provided by RBI, their corresponding grouping by ICAI into Category 2, and the ones added by the author for this study are detailed in Appendix 1.

v. Validation of analysis-Frequency based

We assess whether the mapping dictionary developed is robust enough to identify EWS in all cases and whether there are additional words that need to be added to our dictionary. Cases which yield less than three EWS are analysed manually to identify additional words that are added to the Dictionary. The final Dictionary list contains 1155 words. The comprehensive list is provided in Appendix 1 taking into account Institute of Chartered Accountant of India and RBI's classification and authors own EWS collected during the training of Dictionary.

vi. Validation of analysis-Human audit

We enlist a group of post graduate student volunteers familiar with bank frauds and early warning signals in line with the RBI Master Circular. The students were asked to independently read through 10 cases each and list out all the early warning signals, along with the words in the case that guided them to a particular EWS.

The validation exercise was carried out using counts, hence if a particular EWS appeared more than once in a case it was counted once for the specific case. We compare the count of EWS identified by volunteers with results of the NLP program. The summary of match between our results and those by the independent validation group is given below in Table 3.

The summary indicates that the overall number of EWS identified by the NLP program (Study Results) is 247 which is higher than the number of EWS that have been identified by the human validation test (231). Further, we observe that in 4 out of 6 Major

Category 1 level EWS, the number of EWS identified by the NLP program is more than that identified by the manual validation process. Given this level of match, we believe that our results are robust.

vii. Development of EWS -based ranking/ scoring model

We use two methodologies for developing the ranking and scores for the EWS:

a) First, we see at Category 1 level which are the most important EWS based on their frequencies.

b) Second, we normalize the Category 2 level EWS and rank them.

This frequency-based analysis [presented in Section IV (3)] tells us which EWS appear more in the analyzed cases and therefore give an idea of which of these EWS are important at Category 1 and Category 2 level. The ranking on basis of Normalized scores in presented in Section IV (4).

However, the frequency-based analysis till now does not account for the fraud value. To understand this, we do a logistic regression of fraud amount and the Category 1 EWS. This leads to more nuanced understanding of the EWS relevant for higher value frauds. The results are presented in Section IV (5).

4. Results

1. Survey results

We present the results of the survey conducted to study the effectiveness and ease of identification of the EWS provided by RBI in Table 4 (details in Appendix 3). The bankers have ranked Diversion of Funds and Regulatory Concerns as the two set of EWS at Category 1 level as being most effective for detection of Frauds with a mean score of 4.37 on a maximum score of 5 followed by Issues in Primary/Collateral Securities with a mean score of 4.34 on a maximum score of 5. We observe that the ease of identification of these signals is comparatively lower. Survey results show that bankers have stated that they find it easier to identify issues with respect to Operation in accounts and Diversion of Funds, but the level is lower at an average score of 4.20 on 5 and 4.02 on 5. The entire survey results are presented in Appendix 3 for both effectiveness and ease of identification.

2. Description of Data used for NLP program

Our data comprises 653 credit fraud cases extracted from different sources detailed in Table 1. We have tabulated the FIRs by case numbers and collected data on few additional fields of interest as detailed earlier. The additional fields include fraud amount, industry and legal entity of borrower, duration of fraud, nature of banking relationship with respect to the loan account.

We observe that the largest number of fraud cases involve Private Limited companies at 40% of the reported cases in our data set. However, in terms of value, Public Limited Companies account for 75% of the total loss. Similar, results were reported by Mohanty (2020). Next, we examine the industry of borrowers in Table 6. The Manufacturing sector contributes to both the largest number of cases at 37% and largest value of fraud losses at 54% of the total loss suffered by banks. This is in line with Deloitte (2014).

Exploring the nature of the banking relationship on whether consortium or non-consortium accounts contribute to maximum losses, we observe in Table 7 that the number of cases involving Consortium accounts is smaller at 21% of the total cases. However, Consortium relationships account for 67% of the loss in cases examined. We understand that this may be because consortium arrangements are most likely to be present in large value loans.

Chart 1 provides important information relating to magnitude of fraud losses and delay in detection. We observe that the average fraud value per case steadily increases as the total fraud duration or delay in detection increases.

3. Frequency based analysis of EWS

The analysis of 653 files using NLP methodology gives a total of 7176 EWS. Table 8 gives the major Category 1 distribution of EWS. In terms of frequencies, Diversion of Funds, Concealment or Falsification of Documents and Weakness in Internal Controls at banks are the most important EWS, leading to frauds in Indian banks.

In Table 9, we see top ten Category 2 EWS in the total 7176 EWS generated. 'Siphoning and/or misappropriation', 'Fake/Fabricated documents' and 'Bribe/Collusion / Theft' are the most important of the EWS in terms of frequency counts. This table presents the ranking of EWS at the Category 2 level without grouping them by Category 1 level.

4. Scores/ Ranking of Category 2 EWS (Normalized values)

To develop the Category 2 level ranking and scoring of EWS we use the min-max normalized scores (Zheng, 2019, p.318). Since the distribution is not known, normalization is used. We use the following formula for normalization:

$$EWS_{norm,j} = \frac{EWS_{i,j} - EWS_{min,j}}{EWS_{max,j} - EWS_{min,j}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where j refers to the category 2 variable and i to the fraud case, i=1, 2,.....641.

For the normalized score, we drop those cases where EWS is below 3 in the final analysis and use 641 cases. We are assuming that in the cases where EWS is identified is below three, the python code has failed to pick up the relevant words and hence it would not right to include it in the normalized Category 2 list.

Table 10 gives the normalized rankings for top 10 Category 2 EWS. Appendix 4 gives the full list of Category 2 level EWS and their scores.

The results we find are largely similar to the frequency-based analysis and Bribe/collusion/theft, Siphoning and/or misappropriation, Fake/fabricated documents, Frequent change in scope, and False information/ hiding facts to be most important. To compare further, we list in Table 11 the top ten Category 2 EWS in terms of frequencies and on the basis of normalized scores. We see while majorly the results are the same for both methods, some important EWS which are underestimated when we look at frequencies only come up in normalized ranking like RTGS to unrelated parties and Resignation of the key personnel /frequent changes in the management.

Pearson's correlation was run to assess the relationship between the presence of various early warning signals at the Category 1 level. We observe a positive correlation between diversion of funds and inter-group concentration of transactions of 0.35 at a 1% level of significance. Diversion of funds also has a weekly positive coefficient of correlation with operation in accounts of 0.24 and with concealment or falsification of 0.20 at the 1% level of significance. Concealment or falsification has a weekly positive correlation with operation in accounts of 0.20 at the 1% level of significance. Besides, Business failure has a positive correlation of 0.26 with regulatory concerns at the 1% level of significance. Insider trading also has a positive correlation with regulatory concern of 0.23 at the 1% level of significance (Table in Appendix 5).

4. Fraud quantum and EWS: Logistic regression results

To understand better if some of the category 1 EWS variables have different impacts for low and high value fraud, we analyze further using a logistic model. The ordered Logistic model estimates the relation between an ordered and categorical dependent variable (log fraud value) and a set of independent variables as given in equation 2.

$CATV_i$= F (BUISF, FALDOC, DIF, INSTR, INTER, INTFR, COLSEC, OACC, OTHER, REGL, CONS, YRS) ……(2)

where, CATV: Category of Fraud,

BUISF: Business failure,

FALDOC: Concealment or Falsification of documents,

DIF: Diversion of Funds,

INSTR: Insider trading,

INTER: Inter-Group/Concentration of Transactions,

INTFR: Internal Control Weaknesses,

CLOSEC: Issues in Primary/Collateral Security,

OACC: Operations in Accounts,

OTHER: Other Signals,

REGL: Regulatory Concerns,

CONS: Consortium,

YRS: Years.

The ordered categories are estimated as a linear function of independent variables. The dependent variable in our model is CATV, which is categorized into 1, 2, 3, 4 according to the range of log fraud value. We identify from the frequency distribution the four main buckets for the CATV variable. The frequency distribution is given in Table 12a.

The choice of logistic regression comes from the fact that it accommodates different variable types, including categorical variables dealt in here. Again, the frequency distribution with each predictor dependent variable having more than 100 observations allow us to have the 10 variables in accordance to the minimal 10 Events Per Variable rule[5]. Also, we recognize the loss of information from the conversion of continuous to discrete (categorical) variable but the differentiation into high medium and low buckets is crucial for the Logistic analysis used in the paper. To keep the information loss as minimal we increase the number of categories (Pasta, 2009) to 4.

Table 12 b presents the results for the ordered logistic regression. We seek to understand the determinants of the category of fraud (ranging from low value to high value) in terms of the independent variables. The model is statistically significant and shows variables Diversion of Funds (DIF), Inter-Group/Concentration of Transactions (INTER), Issues in Primary/Collateral Security (COLSEC), Consortium (CONS) have statistically significant impact of fraud quantum category. An increase in the presence of the EWS pertaining to each of Category 1 increases probability of quantum of fraud.

The marginal effects help to understand the impact of each variable on fraud quantum category. We see that an increase in number of EWS pertaining to Diversion of Funds, Inter-Group/Concentration of Transactions, Internal Control Weaknesses, Issues in Primary/ Collateral Security, Operations in Accounts, leads to an increase in quantum of fraud. An increase in EWS pertaining to the above-mentioned categories lower the probability of low and medium value frauds and increases the probability of higher value frauds.

An increase in INTER leads to 11% decrease in probability of fraud being in lowest category. An increase in COLSEC leads to 3% decrease in the chance of fraud being in lowest category. Similarly, an increase in DIF leads to 1% decrease in the chance of fraud being in lowest category. Similarly, for CATV being in category 2, we see the increase in EWS pertaining to the major categories (DIF, INTER, INTFR, COLSEC, OACC) leads to a fall in probability of log fraud being in medium range of 1.8 to 3.

We see an increase in EWS increases the probability of higher value frauds (i.e. those in categories 3 and 4). Increase in INTER leads to 15% increase, while increase in COLSEC leads to 4% increase in the chance of fraud being in highest category 4. An increase in DIF leads to 1% increase in the chance of fraud being in highest category 4.

5. Discussions, Policy direction and Limitations

A set of comprehensive early warning signals notwithstanding, detection of fraud is challenging to law enforcement agencies. As financial intermediaries, bankers are trained to understand creditworthiness of their clients but not to detect mala fide intentions of fraud. This is a concern at any point of time, especially so in India today, with the loss frauds growing for the banking sector.

To aid the bankers the mere presence of list of EWS is not enough. The paper develops a ranking and scoring model which differentiates between Early Warning Signals in terms of their importance and presents a grading that can be referred to for understanding fraudulent activities better. The scorecard developed for EWS provides the template for bankers to understand which of the Early Warning Signals contribute

[5] See van Smeden (2016) for a discussion of the 10 EPV criterion and alternatives for smaller data sets

more towards high-value frauds. From the point of view of the regulators, it also spells out wherein more tightening of regulations or detection mechanisms have to be developed to catch fraud at nascent stage.

The survey results point out that the Early Warning Signals are effective but not easy to identify in the opinion of the bankers. The difficulty in identifying EWS as brought out by the survey brings out the importance of ranking EWS by importance so that scarce resources can be focused on identification of the most significant EWS.

For the fraud cases included in the study we find that larger number of frauds are reported with respect to Private Limited Companies, but Public Limited Companies account for a larger percentage of frauds in value terms. Sectoral comparison suggests that Manufacturing Companies account for both a larger number and value of Cases reported followed by Trading Companies. This information can be used by fraud risk team, internal audit teams and supervisors to focus on monitoring for high value fraud. Consortium banking arrangements account for a larger percentage of fraud losses, which is possibly because of larger loan values within purview of Consortiums. However, bank management should relook at Consortium processes and monitoring to ensure that lapses do not occur because of a lack of accountability in a collective responsibility situation. The data indicates that on an average fraud losses per case increases with a delay in reporting. This brings out the importance of setting up controls for early detection of frauds.

This paper finds that Diversion of Funds, Concealment or Falsification of documents, Internal Control Weaknesses at Category 1 level are important in terms of their frequency of occurrences. The presence of Early warning signals from Diversion of Funds (DIF), Inter-Group/Concentration of Transactions (INTER), Issues in Primary/Collateral Security (COLSEC), makes it very likely that frauds would be in the high-value category. Diversion of funds and intergroup activities are closely related and can be important as early warning signals and detection of any of the subcategories within the two product categories should make bankers vigilant about it. Diversion of funds increase the chances of high-value fraud and therefore there has to be more vigilance in this regard. We have also found that issues related to primary collateral are important. This calls for a greater check of title deeds submitted to the bank, including a cross verification if required, stock checking at regular intervals and any discrepancy to be immediately noted by the bank.

Moreover, as seen within Inter-Group/Concentration of Transactions (the most important EWS for high value frauds) the most common early warning signals is substantial related party transactions. Substantial related party transactions can mean the presence of sister concerns or off-shore entities, and in such cases, the information is likely to be available with a bank. In fact, compared to misinformation or misleading information, it would be easier to set internal processes to be more vigilant in case of substantial related party transactions.

The scoring developed in this paper helps us to understand what early warning signals are likely to be more important in major cases of fraud and specially in case of high-value frauds. It helps to underline which signals to have a look out for and which can become Achilles' heel especially for high value frauds. This study provides useful inputs to the regulator with added EWS arrived from the data collected during the study, as also help to tighten loopholes around the more important EWS.

Our study is limited by the fact that it uses the data availed from CBI and not directly from the banks. Moreover, we could not get the data pertaining to frauds where private sector banks are lead bankers. The frequency-based methodology may show overcounting of EWS, if the same EWS appears many times in different ways in the same FIR. While the reports are generally very concise, we cannot rule out the possibility of overcounting completely.

References

Abdul Majid, G.F.A. and Tsui, J.S.L. (2001), An analysis of Hong Kong auditors' perceptions of the importance of selected red flag factors in risk assessment, *Journal of Business Ethics*, 32, 263-74

Albrecht, W.S., Albrecht C.O. Albrecht, C.C. & Zimbelman, M.F(2012). *Fraud Examination*, Cengage Learning, Mason(USA)

Apostolou, B., Hassell, J., Webber, S. and Sumners, G.E. (2001), The relative importance of management fraud risk factors, *Behavioral Research in Accounting*, 13, pp. 1-24.

Bajaj, R. and Krishnakumar, D. (2020), Overhang of NPAs problems, Paper presented at SIMSARC, 2020.

Beasley, M.S., Carcello, J.V., Hermanson, D.R. and Lapides, P.D. (2000), Fraudulent financial reporting: consideration of industry traits and corporate governance mechanisms, Accounting Horizons, 14 , 4, pp.441-54.

Bell, T.B. and Carcello, J.V. (2000), A decision aid for assessing the likelihood of fraudulent financial reporting, *Auditing: A Journal of Practice & Theory*, 19 (1), pp. 169-84.

Burns, S. (1997), "The honourable fraudsters", Accountancy, September, p. 39 as cited in Smith et al. (2005)

Calavita, K.; Henry N P.; & Robert H T., (1997), Big Money Crime: Fraud and politics in the saving and loan crises, p.9.

Coenen, T. L. (2008), *Essentials of Corporate Fraud*, John Wiley & Sons, Hoboken.

Deloitte (2014), Decoding Frauds in Manufacturing Sector, https://www2.deloitte.com/content/dam/Deloitte/in/Documents/finance/Forensic-Sector-Services/in-fa-manufacturing-noexp.pdf

Ghafoor, A., Zainudin, R. & Mahdzan, N.S. Factors Eliciting Corporate Fraud in Emerging Markets: Case of Firms Subject to Enforcement Actions in Malaysia. *J Bus Ethics* **160,** 587–608 (2019). https://doi.org/10.1007/s10551-018-3877-3

Heiman-Hoffman, B.V., Morgan, P.K. and Patton, M.J. (1996), "The warning signs of fraudulent financial reporting", Journal of Accountancy, October, pp. 75-6

ICAI (2018), Early Signals of Frauds in Banking Sector.

India Today, May 18, (2018), https://www.indiatoday.in/magazine/the-big-story/story/20180521-india-public-sector-bank-npa-vijay-mallya-nirav-modi-debt-bailout-1231739-2018-05-10

Jullum, M., Løland, A., Huseby, R.B., Ånonsen, G. and Lorentzen, J. (2020), "Detecting money laundering transactions with machine learning", *Journal of Money Laundering Control*, 23 , 1, pp. 173-186. https://doi.org/10.1108/JMLC-07-2019-0055

Kannan, S and K Somasundaram (2017), "Autoregressive-based outlier algorithm to detect money laundering activities", *Journal of Money Laundering Control*, 20, 2, 190-202.

Koehler, Derek & Brenner, Lyle & Griffin, Dale. (2002). The calibration of expert judgment: Heuristics and biases beyond the laboratory. Heuristics and biases: The psychology of human judgment.

Loebbecke, K.J., Eining, M.M. and Willingham, J. (1989), "Auditors' experience with material irregularities: frequency, nature, and detectability", Auditing: A Journal of Practice & Theory, 9, 1, 1-28

Loughran, T. And Mcdonald, B. (2011), When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66: 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Mohanty, Prasanna (2020), Rebooting Economy XI: Why are private companies so prone to financial frauds?, https://www.businesstoday.in/opinion/columns/indian-economy-why-are-private-companies-prone-to-financial-frauds-corporates-tax-havens/story/411453.html

Nunnally J.C. (1978), An Overview of Psychological Measurement, In: Wolman B.B. (eds) Clinical Diagnosis of Mental Disorders. Springer, Boston, MA. https://doi.org/10.1007/978-1-4684-2490-4_4

Pasta D (2009) Learning when to be discrete: continuous vs. categorical predictors. Paper 248–2009, SAS Global Forum. http://support.sas.com/resources/papers/proceedings09/248-2009.pdf. Accessed Mar 2018

Phua C. Lee, V., Smith, K. & Gayler, R (2010). A comprehensive survey of data mining-based fraud detection research, *Artificial Intelligence Review*, 1–14.

Pincus, V.K. (1989), "The efficacy of a red flags questionnaire for assessing the possibility of fraud", Accounting Organizations and Society, 14, 1/2, 153-63.

RBI (July 03, 2017), Master Directions on Frauds – Classification and Reporting by commercial banks and select FIs https://www.rbi.org.in/Scripts/BS_ViewMasDirections.aspx?id=10477

RBI Annual Report (2019-2020), https://www.rbi.org.in/Scripts/AnnualReportPublications.aspx?year=2020.

Shelton, S.W., Whittington, O.R. and Landsittel, D. (2001), Auditing firms' fraud risk assessment practices, Accounting Horizons, 15,1, 19-33.

Singh K and Best, P.(2019), Anti-Money Laundering: Using data visualization to identify suspicious activity, International Journal of Accounting Information Systems, 34,100418.

Singh, Charan and Pattanayak, Deepanshu and Dixit, Divyesh and Antony, Kiran and Agarwala, Mohit and Kant, Ravi and Mukunda, S and Nayak, Siddharth and Makked, Suryaansh and Singh, Tamanna and Mathur, Vipul, Frauds in the Indian Banking Industry (March 2, 2016). IIM Bangalore Research Paper No. 505, Available at SSRN: https://ssrn.com/abstract=2741013 or http://dx.doi.org/10.2139/ssrn.2741013

Smith, Malcom, Normah, Haji; Omar, Syed; Iskandar; Zulkarnain, Sayd Idris, Ithnahaini Baharuddin, (2005), Auditors' perception of fraud risk indicators, *Managerial Auditing Journal*, 20, 1, 73 – 85.

Stamler, R. Possamai M. & Maraschdorf (2014), *Fraud Prevention and the Red Flays System*, CRC Press, Boca Raton.

Tetlock, P.C., Saar-Tsechansky, M. And Macskassy, S. (2008), *More Than Words: Quantifying Language to Measure Firms' Fundamentals*. The Journal of Finance, 63: 1437-1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

van Smeden, M., de Groot, J.A., Moons, K.G. *et al.* (2016), No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology,* 16, 163 (2016). https://doi.org/10.1186/s12874-016-0267-3

Weisenborn, D. and Norris, D.M. (1997), Red flags of management fraud, NPA, 42 , 2: 29-34

Zheng, Y. (2019). *Urban Computing.* United Kingdom: MIT Press.

Table 1: Sources and number of cases obtained

CBI Branch	*Number of Cases*
Anti-Corruption Branch	317
Banking Security & Fraud Cell	136
Economic Offence wing	169
Special Crime Branch	8
Special Task Force Branch	23
Total Cases	**653**
Sources: Authors	

Table 2: Example of list of words included in Dictionary and EWS mapped to

Case Number/Identifier	*Words in Case FIRs*	*EWS Mapped to*
RC0742019E0005	fictitious entries	annual report discrepant
RC0742018E0003	abnormally high debtors turnover ratio	annual report inconsistency
RC2152017E0004	could not be explained satisfactorily	does not make economic sense
RC2152018E005	fabricated certified copy	fake/fabricated documents
RC0742018E0003	entities could potentially be non-existent	fictitious counterparties or transactions
Source: Authors		

Table 3: Human audit validation report

Major Category 1 EWS	*Number picked up by NLP program*	*Number of EWS given by students*
Concealment or Falsification of documents	73	81
Diversion of Funds	70	49
Inter-Group/Concentration of Transactions	13	12
Internal Fraud	56	27
Issues in Primary/Collateral Security	17	51
Operations in Accounts	18	11
Total	247	231
Source: Author		

Table 4: Survey Results at Category 1 Level

Category 1 EWS	*Effectiveness (Overall Average)*	*Ease of Identification (Overall Average)*
Diversion of Funds	4.37	4.02
Regulatory Concerns	4.37	3.84
Issues in Primary/Collateral Security	4.34	3.90
Operations in Accounts	4.25	4.20
Concealment or Falsification of documents	4.22	3.69
Inter-Group/Concentration of Transactions	4.20	3.73
Other Signals	4.14	3.62
Scoring is on a scale of 1 to 5 with 1 being the least and 5 being the highest score		
Source: Authors		

Table 5: Break-up of accounts by nature of legal entity of borrower

Legal Entity	*No of Cases*	*% Number of Cases*	*Fraud Loss in Rs. Crore*	*% Value of Fraud Loss*
Public Limited	160	25%	69,216.08	75%
Private Limited	261	40%	20,204.20	22%
Sole Proprietorship	82	13%	775.27	1%
Partnership	63	10%	1,148.45	1%
Others	87	13%	1,417.31	2%
	653	100%	92,761.30	100%
Source: Authors				

Table 6: Frauds by Industry of the borrower

Industry of Borrowers	*No of Cases*	*% Number of Cases*	*Fraud Loss in Rs. Crore*	*% Value of Fraud Loss*
Manufacturing	239	37%	50,173.52	54%
Trading	154	24%	12,236.24	13%
Service	61	9%	4,576.99	5%
Construction	43	7%	3,643.03	4%
Agriculture	20	3%	3,161.17	3%
Others	136	21%	18,970.36	20%
	653	100%	92,761.30	100%
Source: Authors				

Table 7: Frauds by Banking Arrangement of the borrower

Nature of Banking Relationship	*No of Cases*	*% Number of Cases*	*Fraud Loss in Rs. Crore*	*% Value of Fraud Loss*
Consortium	139	21%	62298.22	67%
Non-Consortium	514	79%	30463.085	33%
	653	100%	92761.305	100%
Source: Authors				

Table 8: EWS (Category 1 level, by percentage)

Category 1	**Percentage of EWS**
Diversion Of Funds	40.29%
Concealment Or Falsification of Documents	28.93%
Internal Control Weaknesses	18.94%
Operations In Accounts	4.36%
Inter-Group/Concentration of Transactions	3.68%
Issues In Primary/Collateral Security	2.91%
Other Signals	0.65%
Regulatory Concerns	0.13%
Business Failure	0.08%
Insider Trading	0.03%
Source: Authors	

Table 9: Top ten category 2 EWS as percentage of total EWS.

Category 2	**Percentage of EWS**	**Category 1**
Siphoning of funds	32.93%	Diversion of Funds
Fake/fabricated documents	18.94%	Concealment or Falsification of documents
Bribe/collusion/theft	18.16%	Internal Control Weaknesses
False information/ hiding facts	6.56%	Concealment or Falsification of documents
Frequent change in scope	5.30%	Diversion of Funds
Substantial related party transactions	3.65%	Inter-Group/Concentration of Transactions
Delay in outstanding due	2.20%	Operations in Accounts
Fictitious counterparties or transactions	2.08%	Concealment or Falsification of documents
Frequent invocation of BGs and devolvement of LCs	1.55%	Operations in Accounts
Not routing of proceeds through lead bank/lenders	1.24%	Diversion of Funds
Source: Authors		

Table 10: Normalized scores for category 2 EWS (top ten)

Category 2	**Normalized scores**
Bribe/collusion/theft	0.40
Siphoning and/or misappropriation of funds	0.21
Fake/fabricated documents	0.21
Frequent change in scope	0.12
False information/ hiding facts	0.08
RTGS unrelated parties	0.08
Resignation of the key personnel frequent changes in the management	0.07
Substantial related party transactions	0.07
Delay in outstanding due	0.06
Fictitious counterparties or transactions	0.06
Source: Authors	

Table 11: Top 10 Category 2 EWS (Frequency-Based Percentages and Normalized Score)

Rank	*Frequency-Based Percentages*	*Normalized scores*
1	Siphoning and/or misappropriation of funds	Bribe/collusion/theft
2	Fake/fabricated documents	Siphoning of funds
3	Bribe/collusion/theft	Fake/fabricated documents
4	False information/ hiding facts	Frequent change in scope
5	Frequent change in scope	False information/ hiding facts
6	Substantial related party transactions	RTGS unrelated parties
7	Delay in outstanding due	Resignation of the key personnel frequent changes in the management
8	Fictitious counterparties or transactions	Substantial related party transactions
9	Frequent invocation of BGs and devolvement of LCs	Delay in outstanding due
10	Not routing of proceeds through lead bank/lenders	Fictitious counterparties or transactions
Source: Authors		

Table12a: CATV distribution

Bin (log fraud)	Frequency	Cumulative %
1.80	143	23.83%
3.00	159	50.33%
4.20	122	70.67%
More	176	100.00%
Source: Authors		

Table 12b: Ordered logistic regression results

Ordered logistic regression			*Number of observations*	*620*
Wald	chi2(12)	192.52		
Prob	>chi2(12)	0.000		
Log pseudolikelihood	-710.167		Pseudo R2	0.1689
CATV	**Coef.**	**Std. Err**	**z**	**P>z**
CONS	2.14	0.22	9.94	0.000*
FALDOC	0.00	0.03	0.04	0.970
DIF	0.08	0.02	3.42	0.001*
INTER	0.84	0.15	5.43	0.000*
COLSEC	0.21	0.07	2.75	0.006**
OACC	0.11	0.10	1.09	0.275
/cut1	-0.28	0.15	-0.58	0.02
/cut2	1.26	0.17	0.93	1.58
/cut3	2.60	0.20	2.20	2.99

Marginal effects after ologit

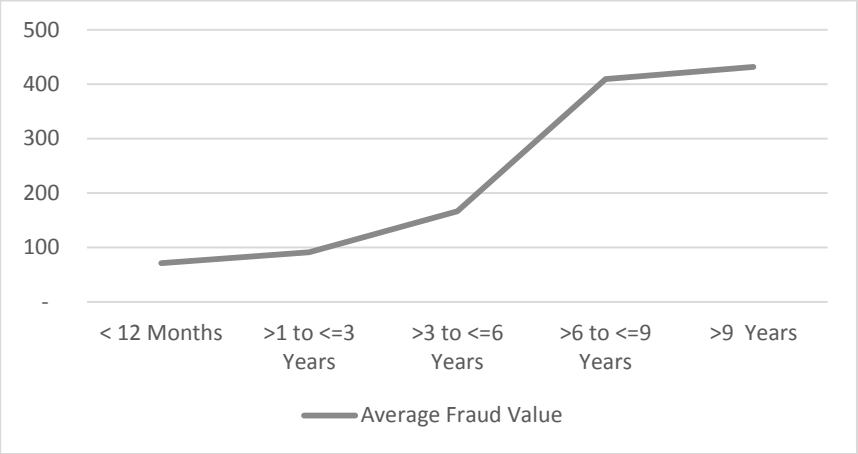
y	=	Pr(catf==1)	(predict,	outcome(1))	
=	0.158164				
variable	dy/dx	Std.Err	z	P>z	
CONS	-0.21	0.02	-10.35	0.000	*
FALDOC	0.00	0.00	-0.04	0.970	
DIF	-0.01	0.00	-3.38	0.001	*
INTER	-0.11	0.02	-5.59	0.000	*
COLSEC	-0.03	0.01	-2.72	0.007	**
OACC	-0.01	0.01	-1.09	0.275	
y	=	Pr(catf==2)	(predict,	outcome(2))	
=	0.308668				
variable	dy/dx	Std.Err	z	P>z	
CONS	-0.24	0.02	-9.74	0.000	*
FALDOC	0.00	0.00	-0.04	0.970	
DIF	-0.01	0.00	-3.23	0.001	*
INTER	-0.10	0.02	-4.58	0.000	*
COLSEC	-0.02	0.01	-2.66	0.008	**
OACC	-0.01	0.01	-1.08	0.278	
y	=	Pr(catf==2)	(predict,	outcome(2))	

=	0.308668				
variable	dy/dx	Std.Err	z	P>z	
CONS	-0.24	0.02	-9.74	0.000	*
FALDOC	0.00	0.00	-0.04	0.970	
DIF	-0.01	0.00	-3.23	0.001	*
INTER	-0.10	0.02	-4.58	0.000	*
COLSEC	-0.02	0.01	-2.66	0.008	*
OACC	-0.01	0.01	-1.08	0.278	
y	=	Pr(catf==3)	(predict,	outcome(3))	
=	0.303222				
variable	dy/dx	Std.Err	z	P>z	
CONS	0.00	0.02	0.19	0.852	
FALDOC	0.00	0.00	0.04	0.970	
DIF	0.01	0.00	2.86	0.004	**
INTER	0.06	0.01	4.19	0.000	*
COLSEC	0.01	0.01	2.36	0.018	**
OACC	0.01	0.01	1.08	0.282	
y	=	Pr(catf==4)	(predict,	outcome(4))	
=	0.229947				
variable	dy/dx	Std.Err	z	P>z	
CONS	0.45	0.05	9.84	0.000	*
FALDOC	0.00	0.00	0.04	0.970	
DIF	0.01	0.00	3.38	0.001	*
INTER	0.15	0.03	4.96	0.000	*
COLSEC	0.04	0.01	2.77	0.006	*
OACC	0.02	0.02	1.09	0.277	

Source: Authors

*where *, **, *** denote statistical significance at 1%,5% and 10% levels respectively.*

Chart 1: Average Fraud value and delay in detection



Appendix 1: List of Early Warning Signals (EWS)

Category 1 EWS	*EWS List as per RBI Master Circular*	*Category 2 EWS identified in the study **
Operations in Accounts	Bouncing of high value cheques	Bounced cheque
	Foreign bills remaining outstanding for a long time and tendency for bills to remain overdue	foreign bill overdue
	Delay observed in payment of outstanding dues	delay in outstanding due
	Frequent invocation of BGs and devolvement of LCs	Frequent invocation of BGs and devolvement of LCs
	Under insured or over insured inventory	inventory under or over insured
	Invoices devoid of TAN and other details	Invoice TAN & detail missing
	Funding of the interest by sanctioning additional facilities	funding interest payment
	Frequent request for general purpose loans.	Frequent General loans
	Frequent ad hoc sanctions	Frequent ad hoc sanctions
	Heavy cash withdrawal in loan accounts	heavy cash withdrawl loan ac
	Significant increase in working capital borrowing as percentage of turnover	significant increase in wc borrowing
Concealment or Falsification of documents	In merchanting trade, import leg not revealed to the bank	import leg not given
	Concealment of certain vital documents like master agreement and insurance coverage	False information/ Hiding facts
	Frequent change in accounting period and/or accounting policies	change in accounting period and/or accounting policies
	Claims not acknowledged as debt high	Claims not acknowledged as debt
	Substantial increase in unbilled revenue year after year	Increase in unbilled revenue
	Material discrepancies in the annual report	Annual report discrepant
	Significant inconsistencies within the annual report (between various sections)	Annual report inconsistency
	Poor disclosure of materially adverse information and no qualification by the statutory auditors	Poor disclosures or No qualification of statutory auditors
		Fake/Fabricated documents*
		Fictitious counterparties or transactions*
		third party transaction/remittance*
		Non-cooperation/ Non compliance *
		Does not make economic sense*
Diversion of Funds	Frequent change in the scope of the project to be undertaken by the borrowers	Frequent change in scope
	Not routing of sales proceeds through consortium /member bank/ lenders to the company	Not routing of proceeds through lead bank/lenders
	High value RTGS payment to unrelated parties	RTGS unrelated parties

Category 1 EWS	*EWS List as per RBI Master Circular*	*Category 2 EWS identified in the study **
	Increase in borrowings, despite huge cash and cash equivalents in the borrower's balance sheet	Increase in borrowings in spite of cash in BS
		Assets not created from loan*
		Siphoning of funds *
Issues in Primary/Collateral Security	Dispute on title of collateral securities	Dispute title collateral securities
	Request received from the borrower to postpone the inspection of the godown for flimsy reasons	Request - postpone godown inspection
	Exclusive collateral charged to a number of lenders without NOC of existing charge holders	Exclusive collateral charged to a number of lenders
	Critical issues highlighted in the stock audit report	Critical issues stock audit report
	Liabilities appearing in ROC search report, not reported by the borrower in its annual report	Liabilities appearing in ROC search report, not reported
	Non production of original bills for verification upon request	Non production of original bills for verification
	Significant movements in inventory, disproportionately differing vis-a-vis change in the turnover	Significant movements in inventory, disproportionately to turnover
	Significant movements in receivables, disproportionately differing vis-à-vis change in the turnover and/or increase in ageing of the receivables	Significant movements in receivables, disproportionately to turnover /receivables
	Increase in Fixed Assets, without corresponding increase in long term sources (when project is implemented)	Increase in Fixed Assets, without corresponding increase in long term sources (when project is implemented)
	Costing of the project which is in wide variance with standard cost of installation of the project	Costing of the project in wide variance with standard cost
		unauthorized removal of stock or less stock*
		Manipulated Stock statements*
		Security valuation*
		Non submission of stock statement*
Inter-Group/Concentration of Transactions	Funds coming from other banks to liquidate the outstanding loan amount unless in normal course	Funds from other bank to pay loan
	Floating front / associate companies by investing borrowed money	Floating front / associate companies by investing borrowed money
	LCs issued for local trade I related party transactions without underlying trade transaction	LCs issued for local trade /related party transactions without underlying trade transaction
	Large number of transactions with inter-connected companies and large outstanding from such companies	Large number of transactions with inter-connected companies and large outstanding from such companies

Category 1 EWS	*EWS List as per RBI Master Circular*	*Category 2 EWS identified in the study **
	Substantial related party transactions	Substantial related party transactions
Regulatory Concerns	Default in undisputed payment to the statutory bodies as declared in the Annual report	Default in undisputed payment to the statutory bodies
	Raid by Income tax /sales tax/ central excise duty officials	Raid by Income tax /sales tax/ central excise duty officials
		Annual returns not filed*
Other Signals	Disproportionate change in other current assets	Disproportionate change in other current assets
	Resignation of the key personnel and frequent changes in the management	Resignation of the key personnel, frequent changes in the management
	Significant reduction in the stake of promoter /director or increase in the encumbered shares of promoter/director	Significant reduction in stake of promoter /director or increase in encumbered shares of promoter/director
		Management failure*
Business Failure		Business Failure*
Internal Control Weaknesses		Bank norms flouted*
		Bribe/Collusion/Theft*
insider trading		insider trading*
*Added by the authors for EWS mapping to the terms use in the CBI documents		

Appendix 2a: Python Code for converting FIR files to text using optical Character Recognition

```
# Import libraries
from PIL import Image
import pytesseract
import sys
from pdf2image
import convert_from_path
import os
```

In [2]:

```
pytesseract.pytesseract.tesseract_cmd = 'C:\\Program Files (x86)\\Tesseract-OCR\\tesseract.exe'
# Path of the input folder
location = "./input/"
for r,d,files in os.walk(location):
   for file in files:
      print("Processing file {}".format(file))
      # PDF_file = "Kwality_false sales and stock_RC2232020A0005.pdf"
      if ".pdf" in file:
         # Creating a text file to write the output
         outfile = "./output/" + file[:-3] + "txt"
         # Open the file in append mode so that
         # All contents of all images are added to the same file
         out = open(outfile, "a")
         '''
         Part #1 : Converting PDF to images
         '''
         # Store all the pages of the PDF in a variable
         pdf_file = "./input/" + file
         pages = convert_from_path(pdf_path=pdf_file, poppler_path= "C:/Users/Dipa/poppler/poppler-
20.11.0/bin", timeout=500)
 page_counter = 1
         # Iterate through all the pages stored above
         for page in pages:
            # Save the image of the page in system
            page.save("page", 'JPEG')
            '''
            Part #2 - Recognizing text from the images using OCR
            '''
```

```
            # Recognize the text as string in image using pytesserct
             text = str(((pytesseract.image_to_string(Image.open("page")))))
            # The recognized text is stored in variable text
            # Any string processing may be applied on text
            # Here, basic formatting has been done:
            # In many PDFs, at line ending, if a word can't
            # be written fully, a 'hyphen' is added.
            # The rest of the word is written in the next line
            # Eg: This is a sample text this word here GeeksF-
            # orGeeks is half on first line, remaining on next.
            # To remove this, we replace every '-\n' to ''.
            text = text.replace('-\n', '')
            # Finally, write the processed text to the file.
            out.write(text)
            print ("Completed processing page {} of {}".format(page_counter, file))
            page_counter += 1
        # Close the file after writing all the text.
        out.close()
        print ("Completed processing {}".format(file))
print("Completed processing all files")
```

Appendix 2 b: Python Code for extracting words from text files and mapping with EWS

```
import nltk
from nltk.tokenize import RegexpTokenizer
from nltk.stem import WordNetLemmatizer,PorterStemmer
import re
lemmatizer = WordNetLemmatizer()
stemmer = PorterStemmer()
import string
from collections import Counter
import os
import numpy as np
import pandas as pd
from os import walk
input_files_path = ***/input_aml_all'
output_file = ***/output'
redflag_file = '***/redflag3.txt'
output_df = None
words_and_flags = []
def setup():
    # read words and flags
    output_columns = ["Filename"]
    global words_and_flags
    words_and_flags = []
        with open(redflag_file, 'r', encoding='cp1252') as file:
        for line in file:
            clear_line = line.replace("\n", '').replace(",", '').replace("'", '').strip()
            word, flag = clear_line.split(':')
            if flag.endswith (';'):
                flag = flag[:-1]
            words_and_flags.append ((word, flag))
            if flag not in output_columns:
                output_columns.append (flag)
    # create output dataframe
    global output_df
    output_df = pd.DataFrame (columns = output_columns)
In [2]:
# function to get words from a file
def get_words_from_file (filename):
    text= open(filename, encoding='cp850').read()
```

```
    sentence = text.lower().replace('{html}',"")
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, '', sentence)
    rem_url=re.sub(r'http\S+', '',cleantext)
    rem_num = re.sub('[0-9]+', '', rem_url)
    tokenizer = RegexpTokenizer(r'\w+')
    tokens = tokenizer.tokenize(rem_num)
    filtered_words = [w for w in tokens if len(w) > 2 ]
    return filtered_words
In [3]:
# get n-grams
def get_ngrams (words, n):
    ngrams_list = list(nltk.ngrams(filtered_words, n))
    ngrams = []
    for ngram in ngrams_list:
        ngrams.append((''.join([w + ' ' for w in ngram])).strip())
    return ngrams
In [4]:
#initialize
setup()
_, _, filenames = next (walk (input_files_path))
file_count = 0
# array of flags
for input_file in filenames:
    flags_list = []
    filtered_words = get_words_from_file (os.path.join(input_files_path, input_file))
    # array of ngrams
    # element 1 is the list of filtered words
    # elements 2-6 are n-grams (2-6)
    ngrams = [None, filtered_words]
    for word, flag in words_and_flags:
        if word in ngrams[1]:
            flags_list.append(flag)
    for n in range (2,9):
        ngrams.append (get_ngrams (filtered_words, n))
        for word, flag in words_and_flags:
            if word in ngrams[n]:
                flags_list.append(flag)
    # set filename column to name of file
```

```
output_df.loc[file_count, "Filename"] = input_file
   # initialize all flag counts to 0
for j in range(1, len(output_df.columns)):
   output_df.loc[file_count, output_df.columns[j]] = 0
   flags_counter = Counter (flags_list)
# set flag column to count of flags
for flag in flags_counter:
   output_df.loc[file_count, flag] = flags_counter[flag]
      file_count = file_count + 1
   output_df.to_excel ('***/aml_result_final.xlsx')
```

Appendix 3: Results from survey of bankers

Rank	*Early Warning Signals*	*Effectiveness*	*Ease of identification*
1	Heavy cash withdrawal in loan accounts	4.62	4.35
2	Not routing of sales proceeds through bank	4.59	4.44
3	Non submission of original bills	4.57	4.35
4	Dispute on title of the collateral securities	4.52	3.79
5	Default in payment to the banks/ sundry debtors and other statutory bodies	4.46	4.08
5	Foreign bills remaining outstanding for a long time and tendency for bills to remain overdue	4.46	4.10
5	Critical issues highlighted in the stock audit report	4.46	4.00
8	High value RTGS payment to unrelated parties	4.44	4.10
9	Frequent invocation of BGs and devolvement of LCs	4.43	4.33
10	Large number of transactions with inter-connected companies and large outstanding from such companies	4.38	3.75
11	Same collateral charged to a number of lenders	4.35	3.96
11	Poor disclosure of materially adverse information and no qualification by the statutory auditors	4.35	3.50
13	Liabilities appearing in ROC search report, not reported by the borrower in its annual report	4.32	3.96
14	Costing of the project which is in wide variance with standard cost of installation of the project	4.30	3.58
14	Concealment of certain vital documents like master agreement, insurance coverage	4.30	3.60
14	Significant movements in receivables, disproportionately higher than the growth in turnover and/or increase in ageing of the receivables	4.30	3.75
17	Request received from the borrower to postpone the inspection of the godown for flimsy reasons	4.29	4.10
17	Increase in borrowings, despite huge cash and cash equivalents in the borrower's balance sheet	4.29	4.00
17	Frequent change in accounting period and/or accounting policies	4.29	3.88

Rank	*Early Warning Signals*	*Effectiveness*	*Ease of identification*
20	Raid by Income tax /sales tax/ central excise duty officials	4.27	3.60
20	Frequent request for general purpose loans	4.27	4.19
22	Substantial increase in unbilled revenue year after year	4.25	3.69
23	Floating front / associate companies by investing borrowed money	4.24	3.29
23	Significant movements in inventory, disproportionately higher than the growth in turnover	4.24	3.71
23	Movement of an account from one bank to another	4.24	4.13
26	Significant increase in working capital borrowing as percentage of turnover	4.22	4.17
27	Significant inconsistencies within the annual report (between various sections)	4.21	3.58
28	Resignation of the key personnel and frequent changes in the management	4.19	3.56
28	Disproportionate increase in other current assets	4.19	3.71
28	Substantial related party transactions	4.19	3.71
28	Material discrepancies in the annual report	4.19	3.73
32	Delay observed in payment of outstanding dues	4.17	4.25
32	Frequent ad hoc sanctions	4.17	4.25
34	Frequent change in the scope of the project to be undertaken by the borrower	4.16	3.54
34	Funds coming from other banks to liquidate the outstanding loan amount	4.16	3.98
36	In merchanting trade, import leg not revealed to the bank	4.13	3.63
37	Funding of the interest by sanctioning additional facilities	4.11	4.33
38	Invoices devoid of TAN and other details	4.10	4.17
39	Onerous clause in issue of BG/LC/standby letters of credit	4.08	3.90
40	Increase in Fixed Assets, without corresponding increase in turnover (when project is implemented)	4.06	3.83
41	Reduction in the stake of promoter / director	4.05	3.58
42	LCs issued for local trade / related party transactions	4.03	3.92

Rank	*Early Warning Signals*	*Effectiveness*	*Ease of identification*
43	Claims not acknowledged as debt high	4.02	3.90
44	Under insured or over insured inventory	3.98	3.88
45	Financing the unit far away from the branch	3.84	4.08
Results of survey of bankers on effectiveness of EWS and ease of identification of EWS for identification of incidents of fraud in their organizations. Scoring is on a scale of 1 to 5, with1 being the lowest and 5 being the highest score.			

Appendix 4 a : Ranking based on frequencies of all Category 2 EWS

Category 2 EWS	*Score*
Siphoning of funds	0.3293
Fake/fabricated documents	0.1894
Bribe/collusion/theft	0.1816
False information/ hiding facts	0.0656
Frequent change in scope	0.0530
Substantial related party transactions	0.0365
Delay in outstanding due	0.0220
Fictitious counterparties or transactions	0.0208
Frequent invocation of BGs and devolvement of LCs	0.0155
Not routing of proceeds through lead bank/lenders	0.0124
Annual report discrepant	0.0078
Bank norms flouted	0.0078
RTGS to unrelated parties	0.0068
Exclusive collateral charged to a number of lenders	0.0067
Resignation of the key personnel frequent changes in the management	0.0064
Security valuation	0.0059
Unauthorized removal of stock or less stock	0.0054
Manipulated stock statements	0.0053
Dispute on title collateral securities	0.0036
Non-cooperation/ non compliance	0.0028
Heavy cash withdrawal loan ac	0.0026
Bounced cheque	0.0022
Significant movements in receivables disproportionately to turnover /receivables	0.0017
Assets not created from loan	0.0014
Funding interest payment	0.0010
Business failure	0.0008
Poor disclosures or no qualification of statutory auditors	0.0008
Important doc not given	0.0006
In merchanting trade import leg not revealed to the bank	0.0006
Non submission of stock statement	0.0006
Raid by income tax /sales tax/ central excise duty officials	0.0006
Annual report inconsistency	0.0004
Increase in unbilled revenue	0.0004
Annual returns not filed	0.0004
Insider trading	0.0003

Category 2 EWS	*Score*
Default in undisputed payment to the statutory bodies	0.0003
Third party transaction/remittance	0.0001
Large number of transactions with inter-connected companies and large outstanding from such companies	0.0001
Lcs issued for local trade /related party transactions without underlying trade transaction	0.0001
Frequent ad hoc sanctions	0.0001
Significant increase in WC borrowing	0.0001
Management failure	0.0001

Appendix 4 b : Ranking based on normalized scores of all Category 2 EWS

Category 2 EWS	*Score*
Bribe/collusion/theft	0.40216
Siphoning of funds	0.214506
Fake/fabricated documents	0.209722
Frequent change in scope	0.117284
False information/ hiding facts	0.080761
RTGS unrelated parties	0.075617
Resignation of the key personnel frequent changes in the management	0.070988
Substantial related party transactions	0.067387
Delay in outstanding due	0.060957
Fictitious counterparties or transactions	0.057485
Frequent invocation of BGs and devolvement of LCs	0.028549
Not routing of proceeds through lead bank/lenders	0.027469
Security valuation	0.021605
Bank norms flouted	0.021605
Unauthorized removal of stock or less stock	0.020062
Manipulated stock statements	0.019547
Assets not created from loan	0.015432
Dispute on title collateral securities	0.013374
Annual report discrepant	0.012346
Exclusive collateral charged to a number of lenders	0.012346
Funding interest payment	0.010802
Heavy cash withdrawal loan ac	0.009774
Business failure	0.009259
Poor disclosures or no qualification of statutory auditors	0.009259
Bounced cheque	0.00823
Non-cooperation/ non compliance	0.007716
Important doc not given	0.006173
Non submission of stock statement	0.006173
Raid by income tax /sales tax/ central excise duty officials	0.006173
Significant movements in receivables disproportionately to turnover /receivables	0.006173
Annual report inconsistency	0.00463
Annual returns not filed	0.00463
Increase in unbilled revenue	0.00463
Default in undisputed payment to the statutory bodies	0.003086
In merchanting trade import leg not revealed to the bank	0.003086
Insider trading	0.003086

Category 2 EWS	*Score*
Frequent ad hoc sanctions	0.001543
Large number of transactions with inter-connected companies and large outstanding from such companies	0.001543
Lcs issued for local trade /related party transactions without underlying trade transaction	0.001543
Management failure	0.001543
Significant increase in WC borrowing	0.001543
Third party transaction/remittance	0.001543

Appendix 5: Correlation Coefficients at Cat 1 Level

Business Failure	1									
Concealment or Falsification	0.09**	1								
Diversion of Funds	0.05	0.20*	1							
insider trading	-0.01	0.12*	0.19*	1						
Inter-Group/Concentration of Transactions	0.03	0.09**	0.35*	0.11*	1					
Internal Control	-0.06	0.03	-0.10**	-0.01	-0.12*	1				
Issues in Primary/Collateral Security	0.04	0.09**	0.16*	0.18*	-0.01	0.03	1			
Operations in Accounts	0.02	0.20*	0.24*	0.14*	0.12*	0.04	0.04	1		
Other Signals	0.04	0.01	0.08**	0.09**	0.04	0.00	0.11*	0.01	1	
Regulatory Concerns	0.26*	0.13*	0.13*	0.23*	0.07***	0.02	0.05	0.21*	0.12*	1
where *, **, *** denote statistical significance at 1%,5% and 10% levels respectively.										