Detection of Frauds in Financial Reporting

Gouri Gopakumar

Abstract: ***Many researches have been done on annual reports to detect whether it is fraud or not by the analytical and empirical part of the report. Annual reports provide information on a company's activities throughout a year. By analyzing the annual report, we can identify the condition of the company whether it is in crisis or operating perfectly. This research deals with the data that can be obtained from the reports' text to determine the probability of being a fraudulent annual report. The verbal content of the report which determines the linguistic features are being analyzed using natural language processing tools to distinguish fraud financial reports from non-fraud financial reports. A set of 60 annual reports were taken for the study. Out of which 30 annual reports are labelled as fraud and the other 30 is labelled as non-fraud. The set of fraudulent companies were selected on the basis of a reporting case of fraudulency of another company or the same company in any other year of non-reporting of cases. The features are selected using a wrapper method search algorithm. A neural network model of MLP (Multi-Layer Perceptron) algorithm is used to classify the data with an accuracy of 85.1%. Classifiers like SVM (Support Vector Machines), Logistic Regression, Naïve Bayes and Random Forest algorithms were also used to identify the best classifier out of all the algorithms. Performance of all the techniques used in this paper are being analyzed and presented in terms of accuracy, precision, recall, F1 score, TN rate and FN rate.***

Keywords: fraud detection, MLP classifier, linguistic features, neural network model

I. INTRODUCTION

Frauds in financial reporting and manipulation of earnings have recently attracted high profile attention [19]. Over the past decade, a number of big-name frauds have been followed by litigation against auditors because of their alleged incompetence in not identifying the fraud in financial statements. As a result, auditors have suffered loss of revenue and the destruction of their reputations, which is much more important. Identifying the symptoms which caused it be fraudulent is the one of most difficult job if it is on the basis of the nature of financial statements [22]. The symptoms might sometimes occur even if there is no fraudulent data.

Generally Accepted Accounting Principles (GAAP) breaches do not automatically mean the existence of fraud, because GAAP departures may be relevant to the business situation and those departures may have been appropriate and recorded (by Policy Board). It is likely that only a small number of signs can occur when fraud occurs. As fraud symptoms may be attributed to valid causes, the mere existence of symptoms can not automatically lead to fraud inference. The signs of fraud, however, cannot be easily measured. In this study, we analyze the text data of the annual reports and by identifying the linguistic features the fraudulent and non-fraudulent annual reports are being differentiated.

II. LITREATURE REVIEW

In a paper [2] , they have used notes to the financial statements of firms have used the text mining method in R statistical system as the notes used contain quantitative as well as qualitative free format information to analyze the extent would be the required details that will be necessary for proper balance sheet and income statement analysis. The result shows that the companies consider themselves to disclose the information more about their capital structure than their asset structure as the companies are afraid to dispose of the risk which are related to liabilities and asset related ones. And it can be stated that the information provided by companies differ by type of company.

In a paper [3] the novel approach used on financial statements are the text mining tools. The result obtained from this are used by the auditors to take decisions to access the risk of fraud taking place with current and future clients. The annually submitted 10-K public company filings from SEC is used to detect the deception and fraud by identifying the linguistic differences in the filings. This paper mainly combines deception theory and to verify the classification accuracy different classification algorithms were used. And the best result was shown by Naïve Bayes and C4.5 with an accuracy of 67.3% The result shows that by Natural Language Processing it helps in determining the veracity by defining and identifying the textual signs which indicate that the flings are fraud or not. They conclude it to be useful by expanding the deception models in detecting the fraud and protecting the public's investments.

In a paper [4], they have used the fraudulent financial reporting instances that are being reported by the SEC in the AAER,.The input variables include the size of the firm, reputation of the corporate, ratios of profitability, activity, asset, leverage and market value and business situation of this firms. The classification is being done by the classification methods such as Logistic Regression, Bayesian Classifiers, BBN, Decision Table/Naïve Bayes classifier, SVM, simple CART, JRip, and logistic model trees, Neural networks (MLP and voted perceptron) and ensemble methods (Bagging, Random Forests, and AdaboostM1). Using different classification measures and other statistical techniques the performance of the methods is being analyzed and evaluated. A sensitivity analysis is being done on the results with the presence of linguistic variables.

Manuscript received on May 25, 2020.
Revised Manuscript received on June 29, 2020.
Manuscript published on July 30, 2020.
***** Correspondence Author

Gouri Gopakumar, Department of Management, Amrita Viswa Vidyapeetham, Amritapuri Campus, Kollam, Kerala

As a result of the classification accuracy, the best result was shown by BBN and DTNB that were considered to be white box classifiers representing relationships within the dataset to be complex. From the result it is clear that with the prediction performance of BBN later on suggests that the variables can help in predicting the non-fraudulent firms in particular.

The result obtained from this research study can be used in need for government and regulatory bodies that can develop a target to interventions directed to firms which are more prone to financial statements fraud. In a paper [5] , they used the financial statements, notes to financial statements and financial news as a source of data and also searched for financial data including names, specific times, account names and money or percentage of validation for business. Three different experiments for each data source. The first experiment was done on annual reports of the financial statements which were issued by the listed companies of 2006. The second experiment was done on notes to financial statements in annual reports and the third experiment was done on the financial e-news derived from the China Times web site. They mainly used recall and precision rates. The research model used mainly focus on the extraction technique in business evaluation and it achieves the best performance in extraction techniques from different data sources.

In a paper [6], the study uses text mining methods to detect high fraud risk in companies. The research sample included companies which have the two main features mainly the pdf converted to Microsoft word format of the board's report for at least four years. CV and LASSO are the methods used for data mining. In this study, they have used many words that explains the fraud risk index in the reports. The LASSO and CVX methods used in this paper described 277 words as terms with the power to clarify the fraud risk index. As stated earlier, in LASSO regressions the dependent variables are in binary mode(0and1), while in CVX they are not discrete and are continuous instead. LASSO regressions are more accurate It suggests that methods which use dependent variables in binary mode are more useful in identifying the fraud risk. For this they used board's reports rather than the information gained from financial statements.

In a paper [7] , the key objective is to guide those who are responsible for choosing the different type of fraud detection techniques within by assessing the tools which are based on financial, language-based, and non-financial. This study is also based on to provide the correlation value between different methods for fraud detection so that they could choose the right method as well as eliminate those which would due an unnecessary cost or duplicate effort. The sample used is of quarterly and annual reports under Accounting and Auditing Enforcement Release (AAER) at least once. These are used since there are no previous assumptions that define which is fraud and which is not. From these releases, the paper uses a decision tree approach in order to distinguish between the fraudulent and nonfraudulent reports. Based on the 200 words thus selected from the list using SVM and Humphreys, Moffit, Burns, Burgoon, and Felix the status of each report is being identified and assigned as the probability value of that report being fraud. This approach doesn't require a pre knowledge about which are the suspicious words and can be easily updated in case if new words are found out from other reports. The accuracy rate of the classification technique used is found out to be 82 percent. By examining the eight different alternatives the F-score value by each method defines the extent of which is an effective method to identify it as fraud. Financial indicators can be useful in helping to detect suspicious firms but instead fail to determine the exact timing of any fraudulent activity within this collection.

In a paper [8], for expressing an opinion by the auditor to identify whether the financial statements are prepared in the applicable financial reporting framework. In this study, they present a context-based machine learning system recommender toolbox that assists the auditor with the tasks of ensuring that the financial statements are accurate. Their key objective was to provide solutions which improve the speed and efficiency of the audit process. The recommender system will have to specify the basic requirement set which the document needs to ensure in it and the document under audit. A logistic regression classifier is used to minimize the logistic loss. A feed-forward neural network model is used to train and minimize the binary cross-entropy. The evaluation metrics for each recommender system is based on the precision and recall value. The system thus created is generic and can easily be tuned and adapted to their checklists ad applications.

In a paper [9], they aim to reveal the bank risk factors from the qualitative textual risk disclosures which are reported in the financial statements. They have used naïve collision algorithm to classify the textual risk disclosures. The textual risk disclosures always contained a summary heading and detailed explanation regarding the risk. Naïve collision algorithm is a rule-based system which has the input obtained by Vector Space Model (VSM) as feature vectors. Using this bank factors, they were able to find the risk condition. The limitation of this particular approach is that it can only consider the text data and not the bank risk factors that contribute it to be a fraud dataset

In a paper [10], the main objective was to find the financial statement outliers deviate from the industries' common practices. A graph is created in similarity metric to calculate the similarity of the financial statements. The technique of clustering is an exploratory device that allows researchers to visually analyse correlations dependent on several variables. Based on the training data collection, a more detailed rule for statistical classification can be established. They affirm the observation that the graph similarity metric is immune to systemic shifts in the balance sheet, and can be used to identify fundamental trends in financial disclosure. This method can induce a class of algorithms developed in text mining and computational biology RNA and DNA sequencing to be applied to the detection of outliers in financial statements. It lays the foundations for finding more significant trends in currency movement expressed in financial statements.

In a paper [11], they have used a text mining approach for identifying the characteristics of the fraudulent reports by analysing the qualitative information. The fraudulent companies are being identified by various sources of information, such as Lexis-Nexis, SEC's Accounting and Auditing Compliance Reports, magazines, the Wall Street Journal etc. the collected textual data of the identified fraudulent companies and fraud detection algorithm such as decision tree algorithm, Help Vector Machines, Naïve Bayes on text information removed.

A performance evaluation measures like precision and recall is done on the algorithm.

In a paper [12], they aim to use text mining to analyse the internal audit role of weapons, in particular the functions of planning and reporting accurate financial statements. They reviewed Greek public corporations' financial reports listed on the Athens stock exchange.

Text Mining techniques, are used to create the term-document matrix containing the TF-IDF values for the terms chosen. After the final dataset was designed, they created the model for assessing the internal audit feature. The Linear Regression model is used for this purpose. The findings show that there is a close connection between the content of the internal audit and the keywords picked. In this study unigrams are used. If by changing it into N grams more continuous sequence of words could be captured to yield a much more improved result. Such keywords could re-evaluate certain specifications and conditions arising from regulatory regimes, such as the US SOX.

In a paper [13] they offer a new standard technique for deviation detection which has a standard CG implemented with embedded synonyms and a set-based dissimilarity function. This method shows comparable results, provided the same sentences, with expert rating. They have followed a hybrid approach that incorporates the rule-based method. This method avoids the rigidity of matching exact phrases, so it is ideal for matching sentences with different terms that has similar meanings. It is considered highly suitable for massive text repositories.

In a paper [14], Conceptual graphs are being used on a Islamic Bank dataset. The sentences obtained are parse and transformed into CGIF. The goal of the experiment is to assess the effectiveness of the proposed algorithm. The result was a line graph. While the resulting external sentences were recorded in a tabular form and showed why they were classified as external. In addition to using a graph, they also measured the effectiveness of their system and CG-dice, associated with human judgements using correlation coefficients. One of the most important contribution of the work is that the synonym specifically incorporated in conceptual graphs. This allows for verbal sentence matching.

In a paper [15],they examine text analytics and information retrieval literature as important for the accounting domain. Current work focuses on a limited range of linguistic computational characteristics. The Internet is also a significant source of electronic computerized analytical papers. In addition to the annual report, IPO prospectus, and press releases, each big organization maintains a web site that includes other documents. These documents can also be used to extract information of value for economic decision-making process. Information economy researchers may apply text analytics and information retrieval techniques to prediction studies of earnings and returns, as well as studies of economic events such as fraud, bankruptcy, and mergers and acquisitions.

In a paper [16], they have highlighted the six areas as the Hinton's paper on Deep Learning which seek to learn representations of words that closely reflect their syntactic and sematic nature. They can be found in language models, retrieval of information, machine translation document classification, and sentiment analysis., Blei's survey on Topic Models it defines the main objective of text mining is to understand the framework that is behind a collection of documents. This structure is used in subject models as a collection of topics depending on the language.. It is an unsupervised process which doesn't define the number and nature of the topics. It has to do more with clustering of document than the categorization of documents. The Latent Dirichlet Allocation (LDA) is the central algorithm in subject modelling. This is better understood as a method of generativity: topics reflect a word distribution. To produce the words in a text, one selects a topic first according to the distribution, then a word from that topic. Aggarwal's Introduction to graphical methods which is a graphical way to capture the hidden structure in document classification. The graphical representation of documents has words as their nodes and the connection between them define the words ordering.Deep learning approaches are used to evaluate sentiments, and mostly work with unlabelled data. The overall trend is towards using richer representations integrating more structure and context of the mined texts. The reason for this was that both NLP methods and Text Mining methods did not scale and were unable to handle real-world data. Another field emerging is the integration of text mining into complex structures involving multimodal data.

In a paper [17], the analysis was to evaluate the CSR report with the keyword extraction algorithm by text mining. The. The CSR reports released in PDF format were converted to text file for decoding it according to the requirements. Using the correspondence analysis, the variability of concepts and variables to find the hidden terms that contribute towards decision making. The result shows that for CSR study, violence terms are chosen for sampling companies to display their success in the environmental context rather than constructive ones. Perhaps the motivating terms are a hypothetical interpretation, variant or expectation without a specific purpose, the distinctive teams are linked to each other and with certain topics, concentrating on the environmental aspect of the CSR survey. In fact, the concept layout is combined with the hierarchical network and occurs alongside deeper connection. They physically detect the distinctive words of a particular motif from the built layout, and the resemblance to the pattern. It provided essential proof of follow-up while carrying out similar activities.

In a paper [18], , an imprint such as deceit was found to be detectable using extractable linguistic structures using natural language processing methodsThe best performing classifiers are being identified by the metrices such as Kappa which compares the observed accuracy with expected accuracy, Accuracy which gives the number of correct predictions, Sensitivity, Specificity, No information rate, P value, Pos Pred value which is the percent of positive fraud values, Neg Pred Value which is the percent of non-fraud values, and Balanced Accuracy which is the arithmetic mean of sensitivity and specificity values. SVM and SGB demonstrate the highest overall to reveal the fraudulent dataset out of the whole.. SVMs are usually best at classifying text because it is characterized by high dimensional space. This is because of their ability to isolate the details from specific boundaries. The worst performing classifier was logistic regression, suggesting that the results are not linearly separable. Further usage should be made of natural language processing methods that identify more nuanced variations in narrative sentiments.

III. METHODOLOGY

Data Description

The list of fraudulent companies use in this study is charged with fraudulent financial-social reported between 2017 and 2019. The companies which were alleged to have violated the Foreign Corrupt Practice Act are also taken into the fraud labelled dataset. [39]

The fraudulent companies were identified according to the risk-criteria which were internally developed as well according to the SEC priorities and these are utilized by the experts of the other divisions of SEC to identify the fraudulent activities done by the companies. They use a more proactive method to identify the activities to conduct investigations with improved effectiveness and efficiency.

A sample of 30 fraud companies from the period between 2017 and 2019 were selected for analysis. During this two-year period 49 companies were identified as fraudulent under the FCPA act and of whose 10-K files are taken from the SEC filings. Set of 30 non-fraudulent companies were selected with the other set to make the predictions. The set of fraudulent companies were selected on the basis of a reporting case of fraudulency of another company or the same company in any other year of non-reporting of cases.

Feature extraction

In this study, the fraudulent and non-fraudulent financial statements are taken from the U.S. Securities and Exchange Commissions (SEC) government site. Distinguished fraud financial statements as recorded by the SEC government agency is labelled as fraud and the other statements as non-fraud. The document extracted from the SEC site is downloaded and the text extracted is cleaned, followed by the basic pre-processing of the documents including the text is done. The study's pre-processing step converts the original textual data into a ready-made data mining structure in which the most important text-functions are defined to distinguish between text-categories. It is the process of inserting a new document into an information retrieval framework [23]. The purpose behind pre-processing is to view each document as a vector of features, that is, to break the text into individual words. The methods that used to pre-process the dataset are as follows:

Lower case transformation: Lower casing the text data, while often ignored, is one of the easiest and most effective method of pre-processing text. It applies to most text mining and NLP concerns, which can aid in situations where the dataset is not very big which helps greatly with planned performance quality [40]

Punctuation Removal: Punctuations doesn't provide any extra details when processing text data [40]. Deleting all instances of the punctuations can reduce the training data size

Stop-word Removal: Some of English's most commonly used expressions are obsolete in Information Retrieval (IR) and data mining which are functional language-specific phrases, regular phrases that do not contain knowledge (i.e., pronouns, prepositions, conjunctions) known as Stop-word which is removed and proved to be very important [40].

By the model of bag of words, the text is being represented as the bag of words which disregards grammar and the multiple words. It is the form text representation with numbers. A sentence can be represented using a string of numbers. The vocabulary of the text is being identified and the occurrence of each word is being noted [40].

Feature Selection:

The feature is selected by using the Wrapper method which is a greedy search algorithm that possibly combines all the features and select that combination which could produce the best result for a particular machine learning algorithm [36]. The most optimal feature is selected by using the Sequential Feature Selector function and the evaluation criteria being used is the ROC-AUC criteria. With fit method on the feature selector it is passed to the training and test sets. With the selected features the classification methods are being carried out and worked with.

Splitting the dataset: training and test split*:*

The classification technique/ method is being done in the ratio 70:30. As the training set is more the model finds a better solution. As the training dataset number gets lower. The model will not be able to learn the general principle and will have a bad validation set performance. The performance is checked up by different ratio sets, and the best performance out of the set is finalized to be 70,30 since the output obtained out of the ratios seems to be less.

Classification

Natural Language Processing (NLP) techniques are being used as the methodology in this paper. NLP deals with the study, comprehension and language generation. It involves study of syntax, morphology, semantic analysis, and phonology. In this study, a fraudulent detection model by MLP (Multi-Layer Perceptron) is used in conclusion with different techniques like MLP (Multi-Layer Perceptron), Naïve Bayes Classifier, SVM (Support Vector Machines), Random Forest, Logistic Regression.

MLP (Multi-Layer Perceptron)

MLP (Multi-Layer Perceptron) is a supervised learning algorithm which utilizes a nonlinear activation or logistic function to let it classify the test data [35]. Out of which an MLP classifier implements an MLP algorithm that trains using the process of Backpropagation. It holds the training sample of financial statements to classify it into the known class labels of fraud or non-fraud accordingly with the model [35]. This method uses neurons as the deciding units and to calculate the contribution of each neuron for decision making in the previous layer and the underlying result at the current neuron. This method is based on pattern recognition Already identified fraud and non-fraud labelled companies are fed as input to the network, as supervised where the outcome can be identified as the expected output is given as result. [41]

The MLP networks are always composed of multiple layers. The equation for the MLP network can be represented as:

$$f(x) = f(5)(f(4)\left(f(3)\left(f(2)(f(1)(x)))\right)\right) \qquad (1)$$

And each layer is represented as:

$$y = f(WxT + b) \qquad (2)$$

Where f is the activation function, W is the set of parameters, or weights in a layer, x is the input vector and b is the bias vector. When all the layers are fully connected, the parameters of each unit are independent of all the other units in each layer. And each unit has a unique set of weight assigned to. The input layers used where two and the number of hidden layers being of 20 hidden units and 5 hidden layers. sklearn.neural_network is being used as the package for MLP classifier in this model.

The activation function by default used is relu, which is the rectified linear unit function which returns the function as:

$$f(x) = \max(0, x) \tag{3}$$

MLP has been used for text classification in a paper[24] they have explained about the use of multi-layered perceptron used for financial classification problem.

With improved learning algorithms the financial classification became much more improved with their improvement on accuracies.

In a paper [42] which is mostly on the topic which deals with the models obtained from training on the increasing amount of title training data which is compared to the models from the training on a constant number of full-texts using MLP, CNN and LSTM. The MLP seems to have benefited the modt out of it which outperforms the full-text method.

Naïve Bayes Classifier

The probabilistic classifiers which uses probabilities of train data with which the test data appears to return a reasonable estimate of fraud financial statements. Out of which a Multinomial NB classifier classifies a document based on the number or relative frequency of the word in a document [22]. The Multinomial NB classifier is being used for the particular research study. A provided example defined by its feature vector is assigned the most probable class by the NB classifier [22]. The Naïve Bayes Classification is applied to categorize the document.

Naïve Bayes Classifier is based on the Bayes' Theorem which is stated mathematically as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{4}$$

The initial probability of A without data training can be referred to as the prior probability P(A) which refers to an established assumption about A. Likewise, P(B) refers to the prior probability of the training set of data of B. P(B|A) is the probability of the appearance of the data B with respect to the establishment of A. The aim is to calculate P(A|B) which is the posterior probability of the appearance of data A with the occurrence of B.

In a paper [25] a method of linguistic classification is used based on positive, negative and neutral sentiments written in English and Vietnam. A naïve Bayes classifier in conjunction with sentiment lexicon dictionary is used as the information mining algorithm to prove viability and data retrieval with an accuracy of 98.2%.

In a paper [26] a Naïve Bayes classifier is used for text classification in classifying the documents of online news into the predefined classes such as business, sports, entertainment, political and health and is being done in Visual Basic 2010 in C# language. They have created an automatic text classifier rather than a manual one with an improved accuracy.

In a paper [27] a text classification model using Naïve Bayes and KNN is used to highlight the efficiency and accuracy of these classifiers using the Student Data Set Rapid Miner.

SVM

Support Vector Machines SVM, is a supervised machine learning technique which is based on statistical theory of learning which classifies objects into predefined categories [38]. The SVM is trained to recognize fake transactions by reviewing the labelled financial statements for fraud and non-fraud. SVMs evaluate a hyperplane that better distinguishes positive from negative examples. We extract the features required and split them into training and testing data. After extracting the features from the labelled test dataset, the values are predicted [38].

In SVM, a hyperplane is being used to select the points in the input variable class either to be 0 or 1. And can be explained by the equation:

$$B0 + (B1 * X1) + (B2 * X2) = 0 \tag{5}$$

Where the coefficients B1 and B2 determines the slop of the line and B0 the intercept being found by the learning algorithm and X1 and X2 being the two input variables. The classification is being done based on this line, If the equation gives a value greater than 0 then the point belongs to the first class and if the equation gives a value less than 0 then the point belongs to the second class. A point that lies close to the line is difficult to classify.

Since the classification is done in polynomial kernal SVM the equation can be defined as:

$$K(x, xi) = 1 + sum(x * xi)^d \tag{6}$$

Where x is input and xi being each support vector, the degree of the polynomial is specified as d.

In a paper [43] they used the combined data set of BBC news articles and 20 news groups and used SVM classifier to classify the content into different categories as Atheism, Business, Car, Entertainment, Politics, Space, Sport and Technology, The training accuracy obtained was 96.40% and the overall accuracy being 89.70%. It also gives the view about the factors that mainly affect the classification techniques and its implementations.

In a paper [28] SVM is used to develop a text classification to classify the Indonesian textual information on the web. SVM performs the best among all the other classifiers being used in the paper with an accuracy of 92.5%.

In a paper [44] an SVM algorithm is being used for document classification with the application data set of diabetics, heart, shuttle and satellite data. Different kernel functions of SVM such as linear, polynomial, sigmoid and RBF is being compared in this study. And the best result obtained for is with RBF for infinite data and multi class.

Random Forest

It is carried out which ensemble learning method for classification, regression and other tasks that operate at the training time by building a multitude of decision trees on the fraudulent and non-fraudulent test data collection and giving the fraud or non-fraud class as output which is the mean forecast of the individual trees or class mode . During the training of datasets, each tree in a random forest learns from a random sample set of the given data set points . The predictions are made by averaging the predictions made from each separate decision tree. [44]

Random forest method creates many individual decision trees during training. The final predictions made from the predictions which are pooled together. As they are using the combination of all the prediction results of the trees they can be known as ensemble technique. Feature importance is being measured by calculating the node probability as:

$$\frac{number\ of\ samples\ that\ reach\ the\ node}{total\ number\ of\ samples} \tag{7}$$

For each of the decision tree the Scikit-learn package calculates the node importance by using Gini importance factor which is illustrated by the equation:

$$ni_k = w_k C_k - w_{left(k)} C_{left(k)} - w_{right(k)} C_{right(k)} \qquad (8)$$

Where ni_k is the importance of node k, w_k is the weighted number of samples reaching node k, C_k is the impurity value of node k, left(k) is the child node from left split on node k and right(k) is the child node from right split on node k. And the importance for each of the feature on the decision tree is calculated by:

$$fi_i = \frac{\Sigma_{k:node\ k\ splits\ on\ feature\ i} ni_k}{\Sigma_{l\epsilon\ all\ nodes} ni_l} \qquad (9)$$

Where fi_i is the importance of feature i , ni_k is the importance of node k. This equation can be normalized to a value between 0 and 1 by the equation:

$$norm\ fi_i = \frac{fi_i}{\Sigma_{k\in all\ features} fi_k} \qquad (10)$$

Random forest level is being done by taking te average over all the trees which is calculated by the equation:

$$RFfi_i = \frac{\Sigma_{j\epsilon all\ trees} normfi_{ik}}{T} \qquad (11)$$

Where $RFfi$ is the importance of feature I which is calculated from all the trees in the Random Forest model, $normfi_{ik}$ is the normalized feature importance of I in the tree k and T is the tital number of trees.

In a paper [45] uses Random Forest techniques using N-gram textual feature and visual feature from representative image to examines the problem of classifying news articles. The text is related to news articles which falls under the categories of Business-Finance, Lifestyle-Leisure, Science- Technology and Sports extracted from the news websites of BBC, Reuters, and TheGuardian. The use of both N-gram textual features as well as visual features has led to the accuracy level of 84.4%. Its uses a late fusion strategy rather than the single feature method.

In a paper [31] the random forest algorithm is used for ordering the value of the function and reducing the dimensions. Then, with the random forest algorithm, the selected features are used, and the F-measure values are calculated as weights for each decision tree to build the employee turnover prediction model. The dataset consisting of the employee information of a communication company in China. And the identified key factors are monthly income, overtime, age, distance from home, years at the company and percent of salary increase. And the most important out of it is the monthly income and overtime. This study can be used as an analytical method to identify the employee's turnover.

In a paper [46] it demonstrates the use of text mining with the combination of Random Forest algorithm for the extraction of the critical indicators and news articles related to the stock market movements. It achieved the best accuracy with respect to other classifiers by the news article classification on the basis of bigram features.

Logistic Regression

This method is used when the target to be obtained or the target obtained is categorical. In this study the dependent variable (target) appears to be the identification of fraud and non-fraud financial statements . To predict which class the data set belongs to, a threshold can be set and based on the threshold a probability is estimated and classified into the labelled classes of fraud or non-fraud

The logistic regression is represented by an example equation:

$$y = \frac{e^{(b0+b1*x)}}{(1+e^{(b0+b1*x)})} \qquad (12)$$

Where y is the predicted output, b0 is the intercept and b1 is the coefficient which is for a single input value of x. Every column in the input data has a coefficient b that needs to be learned from the training data.

In a paper [32] a web-based application to classify twitter tweets to the four topics of health, music, sport, and technology using the classification model of Logistic Regression. The trained set was created by using 1800 tweets with 450 each for each topic. An accuracy of 92 % is acquired by using this classification method.

In a paper [33] uses a unigram model and coordinate wise gradient ascent technique with the introduction of logistic regression approach in which the learning can be done with automatic tokenization. A movie genre classification by using the IMDB dataset, book review classification by genre using AMAZON dataset and topic detection for Chinese text using Chinese dataset is being done using this model. Logistic Regression occurred to have the highest accuracy compared to the other classifiers used.

In a paper [34] a logistic regression model is being used to categories Arabic text. This model is being used on Aljazeera Arabic News dataset. For feature selection chi-square is used and a local policy is undergone for building the logistic regression classifier. The news document is being split according to the five classes of Art, Economics, Politics, Science and Sports. When the other classifiers' such as SVM, KNN and GIS precision, recall, and F1 values are compared the most prominent one tends to be the logistic regression model.

Performance evaluation equations:

The performance of a classifier can be calculated using different ways to determine each of its sustainability to measure the success of the learning algorithm. In this research paper Accuracy, FN rate, TN rate. Precision. Recall and F1-score is being used.

Accuracy is expressed as the ratio of the samples correctly classified. The equation for Accuracy is:

$$Accuracy = \frac{TP+TN}{TP + TN+FP+FN} \qquad (13)$$

where TP can be defined as the number of fraud companies classed as fraud, TN can be defined as the number of non-fraudulent companies which is classified as non-fraud.

TN rate which is also known as specificity is the set of companies that are correctly classed as non-fraud which is the percentage of all the non-fraud companies listed.

Equation of TN rate:

$$True\ Negative\ Rate = \frac{TN}{TN+FP} \qquad (14)$$

FN rate is the set of companies that are classed as non-fraud which is the percentage of all the fraud-companies. Equation of False Negative Rate:

$$False\ Negative\ Rate = \frac{FN}{FN+TP} \qquad (15)$$

Precision is the ratio of which shows how the values obtained after classification is close to each other. Equation of Precision:

$$Precision = \frac{TP}{TP+FP} \qquad (16)$$

Recall is that measure which defines the actual fraud companies that were correctly classified. Equation of Recall:

$$Recall = \frac{TP}{TP+FN} \qquad (17)$$

F1 score is the measure that defines the test's accuracy and it tends to be 1 if the classifier has a perfect precision and recall. Equation of F1-Score:

$$F1 - Score = 2 * (\frac{(Precision * Recall)}{Precision+Recall}) \qquad (18)$$

IV. RESULT AND DISCUSSION

The result of the classifiers' accuracy, precision, recall, f1-score, TN rate and FN rate is being shown in table I. Out of which the resultant classifier that is being concluded is Multi-Layer Perceptron with an accuracy rate of 85%

Table I: Classification results

Classifier	*Accuracy*	*Precision*	*Recall*	*F1 Score*	*TN Rate*	*FN Rate*
Random Forest	0.461	0.5	0.43	0.47	0.6	0.2
Logistic Regression	0.5	0.5	0.4	0.44	0.6	0.6
Naïve Bayes	0.7	0.66	0.8	0.72	0.6	0.2
Multi-Layer Perceptron	0.851	0.714	1	0.83	0.8	1.0
SVM	0.6	0.55	1	0.71	0.2	0.0

V. CONCLUSION

Managerial implications and contributions of the study:

In this research, we presented a methodology for the detection of fraud involving linguistic review of annual reports' descriptive material. The study reaches to a conclusion that the textual data of the annual report that provides valuable knowledge for spotting fraud that is not properly identified by financial statistics. The linguistic discrepancies identified in the fraud and non-fraud annual reports are not intended to over-simplify issues related to fraud detection, but rather to provide perspective and clarity into how companies are portrayed, which warrants further inquiry. This study using linguistic analysis has made another dimension to make advances in the fraud detection research and methodology.

Limitations of the study:

This research is often susceptible to several limitations. The data set considered to be fraud in this may not be the ideal dataset that completely documents the evidence of fraud in a financial statement. Companies which are not being identified as fraud are not added to the dataset. And the non-fraud dataset as mentioned above may also contain unidentified fraudulent statements. A limitation of the feature selection method is that it can be computationally be very expensive in case if the feature set obtained is very large. The key drawback of the MLP algorithm is that it cannot promise that, regardless of the way it is trained, the minimum at which it stops during training is global zero. The MLP algorithm can therefore get stuck inside a local minimum. Another limitation of the MLP algorithm is that the user must set the number of Hidden Neurons, setting this value too small will result in the MLP model being underfitted whilst setting the value too high may result in the MLP model being overfitted.

Future Scope of the study:

This model can be used further in Indian Financial Statements. Further study can be done by applying ensembled techniques with MLP to improve the efficiency of the model. This could be done with a large dataset rather than small to get a more accurate classifier for the study. This could further be improved by using gradient boosting frameworks to improve the accuracy of the models currently being used for identifying fraud.

REFERENCES

1. Prasad Seemakurthi, Shuhao Zhang, and Yibing Qi, Detection of Fraudulent Financial Reports with Machine Learning Techniques, 2015 IEEE Systems and Information Engineering Design Symposium, 358-361
2. Fenyves, Veronika & Böcskei, Elvira & Zéman, Zoltán & Tarnoczi, Tibor. (2019). Analysis of the Notes to the Financial Statement Related to Balance Sheet in Case of Hungarian Information-Technology Service Companies. Scientific Annals of Economics and Business. 66. 27-39. 10.2478/saeb-2019-0001.
3. Humpherys, Sean & Moffitt, Kevin & Burns, Mary & Burgoon, Judee & Felix, William. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems. 50. 585-594. 10.1016/j.dss.2010.08.009.
4. Hájek, Petr & Henriques, Roberto. (2017). Mining Corporate Annual Reports for Intelligent Detection of Financial Statement Fraud – A Comparative Study of Machine Learning Methods. Knowledge-Based Systems. 128. 10.1016/j.knosys.2017.05.001.
5. Seng, Jia-Lang and J. T. Lai. "An Intelligent information segmentation approach to extract financial data for business valuation." Expert Syst. Appl. 37 (2010): 6515-6530.
6. Rahrovi Dastjerdi, Alireza & Foroghi, Daruosh & Kiani, Gholam Hossain. (2019). Detecting manager's fraud risk using text analysis: evidence from Iran. Journal of Applied Accounting Research. 20. 154-171. 10.1108/JAAR-01-2018-0016.
7. Purda, Lynnette & Skillicorn, David. (2012). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. Contemporary Accounting Research. 32. 10.2139/ssrn.1670832.

8. Sifa, Rafet & Lübbering, Max & Nütten, Ulrich & Bauckhage, Christian & Warning, Ulrich & Fürst, Benedikt & Khameneh, Tim & Thom, Daniel & Huseynov, Ilgar & Kahlert, Roland & Schlums, Jennifer & Ladi, Anna & Ismail, Hisham & Kliem, Bernd & Loitz, Rüdiger & Pielka, Maren & Ramamurthy, Rajkumar & Hillebrand, Lars & Kirsch, Birgit & Bell, Thiago. (2019). Towards Automated Auditing with Machine Learning. 1-4. 10.1145/3342558.3345421.
9. Wei, Lu & Li, Guowen & Zhu, Xiaoqian & Li, Jianping. (2019). Discovering bank risk factors from financial statements based on a new semi-supervised text mining algorithm. Accounting & Finance. 59. 10.1111/acfi.12453.
10. Yang, Steve & Cogill, Randy. (2011). Balance Sheet Outlier Detection Using a Graph Similarity Algorithm. SSRN Electronic Journal. 10.2139/ssrn.1943613.
11. Bhardwaj, Ms & Gupta, Dr. (2018). Qualitative analysis of financial statements for fraud detection. 318-320. 10.1109/ICACCCN.2018.8748478.
12. Boskou, Georgia & Kirkos, Efstathios & Spathis, Charalambos. (2018). Assessing Internal Audit with Text Mining. Journal of Information & Knowledge Management. 10.1142/S021964921850020X.
13. Kamaruddin, Siti & Hamdan, Abdul & Abu Bakar, Azuraliza & Nor, Fauzias. (2009). Outlier detection in financial statements: A text mining method. WIT Transactions on Information and Communication Technologies. 42. 71-82. 10.2495/DATA090081.
14. Kamaruddin, Siti & Hamdan, Abdul & Abu Bakar, Azuraliza & Nor, Fauzias. (2009). Dissimilarity algorithm on conceptual graphs to mine text outliers. 2009 2nd Conference on Data Mining and Optimization, DMO 2009. 46-52. 10.1109/DMO.2009.5341910.
15. Fisher, Ingrid & Garnsey, Margaret & Goel, Sunita & Tam, Kinsun. (2010). The Role of Text Analytics and Information Retrieval in the Accounting Domain. Journal of Emerging Technologies in Accounting. 7. 1-24. 10.2308/jeta.2010.7.1.1.
16. Indurkhya, Nitin. (2015). Emerging directions in predictive text mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 5. 10.1002/widm.1154.
17. Liu, Sheng-Hung & Chen, Sih-Yu & Li, Sheng-Tun. (2017). Text-Mining Application on CSR Report Analytics: A Study of Petrochemical Industry. 76-81. 10.1109/IIAI-AAI.2017.164.
18. Minhas, Saliha & Hussain, Amir. (2016). From Spin to Swindle: Identifying Falsification in Financial Text. Cognitive Computation. 8. 10.1007/s12559-016-9413-9.
19. Nguyen, Khanh. “Financial Statement Fraud: Motives, Methods, Cases and Detection.” (2010).
20. Samociuk, Martin, Nigel K. Iyer, and Helenne Doody. A Short Guide to Fraud Risk: Fraud Resistance and Detection. Farnham, Surrey: Gower, 2010.
21. Public Company Accounting Oversight Board (PCAOB). (Nov. 15, 2007). Auditing Standards (A4).
22. Goel, Sunita & Gangolly, Jagdish & Faerman, Sue. (2010). Can Linguistic Predictors Detect Fraudulent Financial Filings?. Journal of Emerging Technologies in Accounting. 7. 10.2308/jeta.2010.7.1.25.
23. Katerattanakul, Nitsawan, "A pilot study in an application of text mining to learning system evaluation" (2010). Masters Theses. 4771.
24. Piramuthu, Selwyn & Shaw, Michael & Gentry, James. (1994). A classification approach using multi-layered neural networks. Decision Support Systems. 11. 509-525. 10.1016/0167-9236(94)90022-1.
25. Le, C. C., Prasad, P. W. C., Alsadoon, A., Pham, L., & Elchouemi, A. (2019). Text classification: Naïve bayes classifier with sentiment Lexicon. *IAENG International Journal of Computer Science*, *46*(2), 141-148.
26. Jasneet Kaur, 2 Seema Bhagla (2016). News Classification Using Naïve Baye’s Claassifier IJARCS International Journal of Advanced Research in Computer Science and Software Engineering.
27. Rajeswari R.P, Kavitha Juliet, Dr.Aradhana "Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier". International Journal of Computer Trends and Technology (IJCTT) V43(1):8-12, January 2017. ISSN:2231-2803. www.ijcttjournal.org. Published by Seventh Sense Research Group.
28. Wulandini, Fatimah and Anto Satriyo Nugroho. “Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Diseases.” (2009).
29. Batoul Aljaddouh, Nishith A. Kotak, "Document Text Classification Using Support Vector Machine", International Journal of Engineering Development and Research (IJEDR), ISSN:2321-9939, Volume.8, Issue 1, pp.138-142, January 2020
30. Multidisciplinary Information Retrieval, 2014, Volume 8849 ISBN : 978-3-319-12978-5 Dimitris Liparas, Yaakov HaCohen-Kerner, Anastasia Moumtzidou, Stefanos Vrochidis, Ioannis Kompatsiaris
31. Gao, Xiang & Wen, Junhao & Zhang, Cheng. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. Mathematical Problems in Engineering. 2019. 1-12. 10.1155/2019/4140707.
32. Indra, S. & Wikarsa, Liza & Turang, Rinaldo. (2016). Using logistic regression method to classify tweets into the selected topics. 385-390. 10.1109/ICACSIS.2016.7872727.
33. Ifrim, Georgiana & Bakir, Gökhan & Weikum, Gerhard. (2008). Fast logistic regression for text categorization with variable-length n-grams. Bing Liu, Bing; Sarawagi, Sunita; Li, Ying: KDD 2008 : proceedings of the 14th ACM KDD International Conference on Knowledge Discovery & Data Mining, ACM, 354-362 (2008). 10.1145/1401890.1401936.
34. Tahrawi, Mayy. (2015). Arabic Text Categorization Using Logistic Regression. International Journal of Intelligent Systems and Applications. 7. 71-78. 10.5815/ijisa.2015.06.08.
35. Amin, Muhammad & Ali, Amir. (2017). Application of Multilayer Perceptron (MLP) for Data Mining in Healthcare Operations.
36. Jovic, Alan & Brkić, Karla & Bogunovic, N.. (2015). A review of feature selection methods with applications. 1200-1205. 10.1109/MIPRO.2015.7160458.
37. F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, 2016, pp. 2264-2268
38. Awad, Mariette & Khanna, Rahul. (2015). Support Vector Machines for Classification. 10.1007/978-1-4302-5990-9_3.
39. https://www.sec.gov/spotlight/fcpa/fcpa-cases.shtml]
40. https://machinelearningmastery.com/clean-text-machine-learning-python/
41. https://pathmind.com/wiki/multilayer-perceptron
42. Mai, Florian & Galke, Lukas & Scherp, Ansgar. (2018). Using Deep Learning for Title-Based Semantic Subject Indexing to Reach Competitive Performance to Full-Text. 169-178. 10.1145/3197026.3197039.
43. Batoul Aljaddouh, Nishith A. Kotak, "Document Text Classification Using Support Vector Machine", International Journal of Engineering Development and Research (IJEDR), ISSN:2321-9939, Volume.8, Issue 1, pp.138-142, January 2020
44. Srivastava, Durgesh & Bhambhu, L.. (2010). Data classification using support vector machine. Journal of Theoretical and Applied Information Technology. 12. 1-7.
45. Liparas, Dimitris & HaCohen-Kerner, Yaakov & Moumtzidou, Anastasia & Vrochidis, Stefanos & Kompatsiaris, Ioannis. (2014). News Articles Classification Using Random Forests and Weighted Multimodal Features. LNCS. 8849. 10.1007/978-3-319-12979-2_6.
46. Elagamy, Mazen & Stanier, Clare & Sharp, Bernadette. (2018). Stock market random forest-text mining system mining critical indicators of stock market movements. 1-8. 10.1109/ICNLSP.2018.8374370.

AUTHORS PROFILE

Gouri Gopakumar, 2nd year MBA student at Amrita School of Business, Amritapuri specializing in Operations and Business Analytics. She is a Computer Science Engineering graduate from Kerala University. Her area of interest is in Data Mining and Analysis, Cloud Computing and Big Data, Database Management and Architecture, Data Visualization and Presentation, Quality Management, Supply Chain Management, Service Operations Management and Enterprise Resource Planning. She has featured a visualization term paper on Travancore Dynasty in Tableau (https://public.tableau.com/profile/gouri.gopakumar#!/vizhome/TravancoreDynastyTimelineTermPaper/Dashboard15) ,