# Using machine learning methods to predict financial performance: Does disclosure tone matter?

Mousa, Gehan ; Elamir, Elsayed; Hussainey, Khaled

**International Journal of Disclosure and Governance**

*DOI:*
10.1057/s41310-021-00129-x

Published: 01/03/2022

Peer reviewed version

# Using Machine Learning Methods to Predict Financial Performance: Does Disclosure Tone Matter?

Gehan A. Mousa & Elsayed A.H. Elamir
College of Business Administration, University of Bahrain, Kingdom of Bahrain

Khaled Hussainey
Faculty of Business and Law, University of Portsmouth, Portsmouth, UK

## Abstract

We use three supervised machine learning methods, namely linear discriminant analysis, quadratic discriminant analysis, and random forest, to build models that predict financial performance of sixty-three listed banks from eight emerging markets for 10 years from 2008 to 2017. We use the design science research (DSR) framework to examine whether the textual contents of annual reports in previous years contain value-relevant information for anticipating future performance; thus, these contents can improve the accuracy and quality of predictive models. We combine two groups of variables in the proposed models. The first group is the sentiment analysis of disclosure tone in annual report narratives using the Loughran and McDonald (2011) dictionary, while the second group is the quantitative properties of banks which consist of five variables: firm size, financial leverage, age, market-to-book ratio, and risk. We find that the random forest method provides the best predictive model. We also find that the accuracy and performance of predictive models can be increased by incorporating disclosure tone variables with financial variables. Interestingly, we find that uncertainty is the most important disclosure tone variable. Finally, we find that firm size is the most important variable related to banks' quantitative characteristics. Our study suggests that the analysis of tone through corporate narrative disclosures can be used as a complementary or diagnostic approach rather than an alternative in making decisions by different stakeholders.

**Keywords:** Predictive Models, Financial Performance, Disclosure Tone, Machine Learning, Discriminant Analysis, Random Forest.

## 1. Introduction

The corporate annual report is considered one of the most informative financial communication channels used by stakeholders in making their decisions, and it is used by companies as a communication tool to send messages or create an image that can impact the perception of their stakeholders (Kloptchenko et al., 2002; Magnusson et al., 2005). These reports contain narrative sections and financial statements. Narrative sections include, but not limited to, management discussion and analysis (MD&A), president's letter, and chief executive officer's message Research shows that financial statements lost their relevance (Lev, 1989) and there is an increasing interest in non-financial information, the language or style of writing, and textual structures of these reports (Henry, 2006; Li, 2010; Qiu et al., 2014). For example, Rogers and Grant (1997, p.17) point out that "in total, the narrative portions of the annual report provide almost twice as much information as the basic financial statements". In this regard, the Association for Investment Management and Research (AIMR) in 2000 used a survey to examine the importance of the MD&A section in annual reports. Their findings reveal that 86% of surveyed financial analysts consider the MD&A section to contain value-relevant information for analysts in assessing firm value and for other stakeholders in making their decisions.

Many studies on the prediction of corporate failure or bankruptcy (Altman et al., 2017; Lukason and Laitinen, 2019) as well as firm performance (Chan et al., 2005; Onder and Altintas, 2017) are based on financial ratios. Aziz and Dar (2006) argue that financial ratios provide the best chance for analysts to predict firm failure. However, the use of financial ratios to predict corporate failure or firm performance has been criticised (Cooper et al., 2000; Yu et al. 2014). Some of these criticisms are that small or medium companies may not publish their financial statements. In addition, financial statements may be exposed to manipulation to achieve certain purposes (Ho and Zhu 2004; Yu et al. 2014). In this case, calculation of financial ratios based on these statements might be misleading for decision-makers and the use of these ratios in building prediction models makes these models invalid. Appiah et al. (2015) reviewed 83 studies on predicting corporate failure covering the period from 1966-2012 that use 137 models. The authors point out that "the neglect of non-financial information remains one of the major gaps in the extant corporate failure literature" (Appiah et al., 2015, p.469).

Research also shows that narrative disclosures can be used as indicators to predict future performance and improve the decision-making process (Qiu, 2007; Dias and Matias-Fonseca, 2010; Qiu et al., 2014). Annual report narratives play a critical role in maintaining market efficiency as financial data in financial statements (AIMR, 2000). Several studies document that textual disclosures in annual reports such as the MD&A section, or other disclosures such as press releases, play a vital role in making investment decisions because they contain a stock of information that can be used for different purposes, such as predicting future financial performance (Li, 2010), stock market reaction (Feldman et al., 2010), earnings (Davis et al., 2006; Li, 2006a), or corporate failure (Balcaen and Ooghe, 2006; Appiah et al., 2015). The prediction of financial performance is one of the most important topics to attract the interest of many stakeholders (Qiu (2007; Balakrishnan et al., 2010; Qiu et al., 2014).

Accounting literature explores the impact of narrative disclosure on performance. For example, Davis et al. (2006) investigate the relationship between disclosure tone in earnings press releases using a measure based on counting optimistic/pessimistic words, and future financial performance measured by return on assets (ROA). The authors find that managers use different tones (optimistic and pessimistic) to provide reliable information on future performance to the stock market, which responds significantly to tone usage. Henry (2006) examines the stock market reaction to disclosure tone in press releases. She finds that the incorporation of variables that reflect the nature of the verbal method used in writing for press earnings releases increases the accuracy of prediction of the market reaction. Feldman et al. (2010) show that the changes in disclosure tone (using positive and negative words) in annual reports are linked with stock market reactions – in other words, the more positive the tone, the higher the stock market return. Li (2006a) uses a fog index to measure the readability of textual contents in annual reports to study their relationship with future earnings and the persistence of these earnings. In another study, Li (2006b) tests the relationship between using some words in annual reports, namely ‘risk’ and ‘uncertain’, and future earnings and stock returns. All these studies provide clear evidence that the textual contents of companies’ financial communication channels are useful for predicting corporate performance.

The current study argues that the textual contents of annual reports such as disclosure tone contain valuable information that can improve the quality and accuracy of predictive models of firm performance. We develop predictive models of corporate financial performance using three techniques of machine learning: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and random forest (RF). Three of these models were based on a set of financial and non-financial variables, while the fourth model contained only financial variables. We ran these models using a sample of 63 conventional banks from eight emerging markets (Egypt, Jordan, Bahrain, United Arab Emirates, Saudi Arabia, Kuwait, Qatar, and Oman) from 2008 to 2017.

There is potential practical relevance for our study as different stakeholders are keen to predict corporate future performance. The study offers three important contributions. First, most predictive studies rely on financial ratios or quantitative data, while our study combines qualitative data through sentiment analysis of disclosure tone in banks' annual reports with quantitative data – namely, the properties of banks such as size, financial leverage, age, market-to-book ratio, and risk level. Our study extends the narrative disclosure literature by quantifying tone across corporate narrative disclosures for a sample of banks from emerging markets to use in building a predictive model for banks' future performance. We provide new evidence that the quality of predictive models of financial performance can be improved by incorporating disclosure tone variables in these models. Second, our study methodologically contributes to the literature by building predictive models that use three techniques of machine learning, namely LDA, QDA, and RF. The proposed models can be a useful tool for all stakeholders, especially shareholders and investors, as the models can help to predict corporate financial performance based on the tone of corporate disclosure. Third, it is one of the first studies – to the best of our knowledge – dealing with this vital subject in emerging markets, as most previous studies have been conducted in developed countries. These markets undoubtedly suffer from the scarcity of studies dealing with this subject.

The paper is organised as follows: Section 2 provides an overview of corporate disclosure theories. Section 3 reviews the relevant literature and develops the research hypothesis. Section 4 provides details on the empirical research framework. Section 5 presents the analyses and the findings. Section 6 shows a comparison among prediction models. Section 7 concludes.

## 2. An overview of corporate disclosure theories

A number of theories have been used to explain corporate annual report voluntary disclosure practice such as agency theory, signaling theory, legitimacy theory, impression management theory, and others. Agency theory argues that there is a conflict of interests between managers and corporate owners. Consequently, managers take actions that maximise their benefits. Also, managers may use voluntary disclosure to reduce information asymmetry and agency costs (Jensen and Meckling, 1976). Signaling theory posits that corporate managers use voluntary disclosures to send messages to stakeholders (Connelly et al., 2011). Furthermore, legitimacy theory argues that corporate managers use their voluntary disclosures to legitimise corporate activities (Hahn and Lülfs, 2014).

The 1990s witnessed the emergence of impression management theory, one of the unique theories explaining managers' motivations for voluntary disclosures. Tedeschi and Riess (1981, p.3) define impression management "as any behavior by a person that has the purpose of controlling or manipulating the attributions and impressions formed of that person by others". Impression management theory goes back to Leary and Kowalski (1990), who identify two types of impression that people try to convey of themselves: impression motivation and impression construction. The first type "refers to how motivated people are to control how they are perceived in a particular social encounter, while the second type refers to the particular image a person will try to convey to others" (Leary, 2001, p.7246). Impression management theory posits that managers seek to influence interpretations related to corporate financial reports by affecting the readers' awareness of these reports (Falschlunger et al., 2015). Hackfort et al. (2019) argue that the process of managing impression takes place in the conscious or subconscious, where individuals seek to control and regulate the impressions that others make of them in different situations. Consequently, they can obtain many advantages by creating a positive impression of them. Leary and Kowalski (1990) argue that managers have a motivation to control their impression to affect the way people view them. They may conduct self-serving interests to adopt a specific agenda in financial reports through manipulating the readers' perception (Beatti and Jones, 2000). Rahman (2012) argues that managers may take advantage of corporate annual reports to provide self-service for the company's financial performance and thus, the management of impressions occurs. Furthermore, Merkl-Davies and Brennan (2007, 2011) point out that economic and psychological incentives can be an explanation for engaging in impression management where managers attempt to maximise their benefits and rewards or minimise the penalties. In our study, we test to see if disclosure tone provides useful information for the prediction of financial performance.

## 3. Literature review and hypothesis development

### 3.1 Literature review

Literature has examined the usefulness of narrative disclosure for stakeholders. For instance, it shows that the tone of risk disclosure contains value-relevant information to investors (Campbell et al., 2014), and the stock market reacts to risk disclosure (Elshandidy and Shrives, 2016; Hope et al., 2016). Kravet and Muslu (2013) find a positive relationship between risk disclosure and stock return volatility. Heinle and Smith (2017) find a negative relationship between risk disclosure and cost of capital.  Lu and Chen (2009) provide evidence that, by using decision tree-based mining techniques, and the classification of companies into good/bad information disclosure, investors can accurately make a rational investment decision.

Another interesting stream of literature addresses the association between the textual content of narrative disclosures and corporate financial performance. For instance, Aly et al. (2018) find a positive relationship between narrative disclosure (in terms of good/bad news) and firm performance. Baginski et al. (2018) report that linguistic tone in earnings press releases, on average, has a positive association with corporate future earnings and incremental value of the market. Furthermore, Clatworthy and Jones (2003, 2006) and Arslan-Ayaydin et al. (2016) use disclosure tone to examine the association between impression management and corporate financial performance. The authors find that managers exhibit opportunistic behaviour when presenting financial performance through their voluntary disclosures. In this regard, Davis and Tama-Sweet (2012) find that managers use a more optimistic tone in their disclosure through the earnings press release in contrast to their tone used through disclosure in the MD&A section, because investors interact more with disclosures by earnings press releases. Using a sample of 110 UK chairman statements of financial institutions from 2006-2010, Ressas and Hussainey (2014) investigate the effect of the financial crisis on the levels of disclosure tone in terms of good/bad news. The main finding of their study shows that financial institutions had more bad news during and after the financial crisis. Similarly, Schleicher and Walker (2010) find that managers use a biased tone when the company's financial performance tends to decline through forward-looking narrative disclosures using a sample of UK firms. Finally, Huang et al. (2014) provide evidence on biased tone by managers. They find that managers apply impression management to manipulate investors' perceptions through their tone in earnings press releases when a company needs to make a critical decision such as issuing new securities or mergers and acquisitions.

Other studies consider the style of writing (readability) in narrative disclosures and its relationship with financial performance, such as Magnusson et al. (2005) who find that the change in writing patterns in the quarterly reports and financial ratios in these reports related to changes in the financial performance of the company. In this regard, Li (2008) finds that the annual reports of firms with a low level of earnings are hard to read. Similar results are reported by Lehavy et al. (2011), who find a relationship between analyst behaviour and the readability of annual reports.

Based on the previous discussion, it can be concluded that there is a general agreement on the importance of corporate narrative disclosures to stakeholders. Our study considers this as a starting point to achieve a different research objective than those reported in the previous studies. It seeks to build predictive models by incorporating disclosure tone as non-financial variables with financial variables. Consequently, it is linked with another stream of disclosure tone studies relating to the prediction of corporate performance.

## 3.2 Hypothesis development

Literature shows that disclosure tone can be a valuable tool for the prediction of financial performance. For example, Balakrishnan et al. (2010) examine narrative disclosures in 10-K and 10K-405 filings to identify whether these disclosures contain value-relevant information. They find that narrative disclosures are positively associated with market performance. Similarly, Li (2010) examines the relationship between disclosure tone related to forward-looking information in the MD&A section and financial performance. He finds a positive relationship between the tone of forward-looking disclosure and future earnings. Lee et al. (2010) use the support-vector machine (SVM) method, embodying both quantitative and qualitative information contained in textual contents of financial reports in their proposed model to predict the movements of stock prices. In this regard, Chen et al. (2009) suggest a model to predict the earnings change of the firm by embodying risk information disclosed in the textual parts of corporate financial reports. Their results show that the accuracy of the earnings model is improved.

Moreover, Qiu (2007) investigates the ability of the textual content of corporate annual reports to predict financial performance. To analyse the textual content of these annual reports, the author uses the 31 dictionaries included in DICTION 5 and three measures of financial performance, namely return on equity (ROE), earnings per share (EPS), and the market response measure (stock return). Then, the author develops three models using SVM techniques. He finds that when the textual content of annual reports is incorporated into predictive models, the performance of these models is improved in terms of the accuracy and Kappa statistics measures. Using the quarterly reports of three big communication firms, Kloptchenko et al. (2002) examine the relationship between the style of writing in these reports and the financial performance of communication firms. The authors employ seven financial ratios including profitability, liquidity, and solvency ratios (as quantitative information) and textual contents of the quarterly reports of firms (as qualitative information). The authors find that the changes in textual contents of the quarterly reports or style of writing are associated with firm performance and help to predict future performance. Using the same seven financial ratios used by Kloptchenko et al. (2002), Magnusson et al. (2005) report similar results. Dias and Matias-Fonseca (2010) examine the association between the language contained in annual reports of 14 listed firms (qualitative data) and financial performance of these firms using a group of financial ratios (quantitative data). They find that the textual contents of annual reports reflect three elements. First, they reflect the firm's financial results in the year in general; second, they refer to the events that led to such results; and finally, they indicate the future changes of these results. Qiu et al. (2014) conduct several experiments to build predictive models of firm performance based on their annual reports using supervised learning methods. The authors provide evidence of the ability of corporate annual reports to be used in predicting the firm performance from one year to the next. However, Hildebrandt and Snyder (1981) count the positive and negative words of 24 annual letters which had been sent to stockholders in 1975 and 1977. They find that firms use optimistic or positive words more frequently than negative ones to reflect their financial results. The authors conclude that there is an inconsistency between narrative disclosures and financial performance.

Moreover, Qiu et al. (2006) employ the SVM method to predict financial performance for the next year. The authors provide evidence on incorporating quantitative and qualitative information in predictive models to improve the performance of these models and increase their accuracy more effectively than using only one type of information in building these models. Meanwhile, de Graaff (2017) analyses the textual contents of 150 annual reports using advanced textual analysis methods to predict financial performance. The author reports that the Fuzzy Fingerprints model is the best model regarding its accuracy (0.8333) and Kappa statistic (0.4973). Moreover, Chou et al. (2018) design a model to study the level of consistency between using financial ratios and textual disclosures in annual reports. The authors analyse the tone of textual disclosures as positive or negative, using the K-means method to classify the firm’s financial performance into good or bad. They find that there is no consistency between textual disclosures and financial ratios between countries. For example, Chinese and Taiwanese companies exaggerate in their textual disclosures, while US firms reverse this behaviour.

The current study seeks to examine how disclosure tone can be a source of prediction of banks’ financial performance. It investigates the power of disclosure tone in building predictive models of financial performance. Consequently, we set the following hypothesis:

The quality of predictive models of financial performance can be improved by incorporating disclosure tone variables in these models.

## 4. Empirical research framework

We test to see if the textual contents of annual reports in previous years contain value-relevant information to predict future performance for the following year; thus, these contents can improve the accuracy and quality of predictive models. Different analytical techniques of machine learning have been employed such as discriminant analysis methods with two main techniques (LDA and QDA), in addition to other methods such as random forest (Breiman, 2001a; McLachlan, 2004; Duda et al., 2012; Siqueira, et al., 2017). To achieve the main goal of this study, we run predictive models using three techniques of machine learning (LDA, QDA, and RF).

*Discriminant analysis* is used to determine the class posterior probability for optimal classification. This probability can be measured as follows:

$$Pr(Y = k|X = x) = f_k(x)p_k \Big/ \sum_{l=1}^{K} f_l(x)p_l$$

Where $Pr(Y = k|X = x)$ refers to posterior probability for membership in class $k$, and $k$ is the number of classes, $f_k(x)$ indicates the density of $x$ in class $k$, and $p_k$ the prior probability.

The LDA will capture the linear relationship in the model through:

$$d_k = log\hat{p}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + x^T\hat{\Sigma}^{-1}\hat{\mu}_k$$

Where $x$ is the feature vector that is to be classified, $\hat{p}_k$ is the estimated prior probability of class $k$, $\hat{\mu}_k$ is the estimated mean of class $k$, and $\hat{\Sigma}^{-1}$ is the estimated inverse pooled covariance matrix. The QDA will capture the quadratic relationship in the model through (Hastie et al., 2009, p.108):

$$d_k = \log \hat{p}_k - \frac{1}{2}\log\left[\hat{\Sigma}_k\right] - \frac{1}{2}(x - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1}(x - \hat{\mu}_k)$$

The conditions for the use of LDA and QDA are that the training data should have a multivariate normal distribution and all classes in LDA are assumed to have the same covariance matrices. In QDA, the covariances are assumed to be different in each class (for more details about LDA and QDA, see Sharma, 1995; Hastie et al., 2009, pp.108 and 110).

The random forest is introduced by Ho (1995, 1998) and Breiman (2001a) who explain it as an ensemble learning method that constructs many decision nodes or trees in the training stage and can be used for classification by taking the mode of the categories and regression by taking the average prediction of the individual nodes.

The steps of the random forest for regression and classification are described by Hastie et al. (2009, p.588) as follows:

1. Select the number of runs (B)

    a. From the training data select a bootstrap sample $S^*$ of size $N$,

    b. To bootstrapped sample, expand a random forest tree $(R_b)$ by iterating the next procedures for every branch node of the tree, till the lowest $(n_{min})$ node size is obtained.

        i. Choose $l$ variables randomly out of $k$ variables.

        ii. Select the superior variable/divide-value among the $l$.

        iii. Divide the node into two child nodes.

2. Produce the collection of trees $[R_b]_1^B$.

For a forecast at a new value $x$:

“Regression:” $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} R_b(x)$.

"Classification:" Let $\hat{C}_b(x)$ be the forecasting set of the $b$th random forest tree. Hence, $C_{rt}^B(x) = majority\ vote\ \left[\hat{C}_b(x)\right]_1^B$.

With this algorithm, the random forest improves the dispersion reduction of bagging via reduction of the association between the trees, without increasing the dispersion too much (Breiman, 2001b, 2017; Hastie et al., 2009).

For feature selection, the Boruta algorithm is introduced by Kursa and Rudnicki (2010) to identify the most interesting and important features of a data set that have an impact on the outcome variable. A large set of independent variables can give rise to heavy computational time and a high risk for overfitting in the data. Importantly, the selection of the significant features before running random forest reduces the estimated time and makes the interpretation of the model results easier. From Kursa and Rudnicki (2010, p.3), the Boruta algorithms can be organised in the following steps:

1. Make duplicate copies of all explanatory variables where all variables are not less than five in the original data.
2. Remove the correlation of explanatory variables with the target variable by shuffling the values of added duplicate copies.
3. Mix the shuffled copies with original ones.
4. Carry out a random forest method on the mixed dataset and accomplish a variable importance measure such as "mean decrease accuracy" to show the significance of each variable.
5. Calculate mean divided by the standard deviation of accuracy loss (Z score).
6. Obtain max. Z score among variables (shadow attributes).
7. If the importance of a variable is lower than max. Z score, tag it as insignificant and it can be removed from the model.
8. If the importance of a variable is higher than max. Z score, tag it as significant and it can be retained in the model.
9. For prespecified random forest runs, repeat the above procedures or until the importance of variables is designated for all shadow attributes.

The key procedure in building a predictive model of firm performance when using these techniques is to establish two data sets: one to build a training model and another for testing the model. The evaluation of the predictive model on new data is important, as the accuracy of prediction can be established by considering how the model is performing on new observations that were not included in the fitting model. The testing model should give a good indication of how well the model is likely to predict new data (Hastie et al., 2009).

Our research framework in the current study is based on design science research (DSR) as a new approach (Horváth, 2007; Hevner and Chatterjee, 2010; Gregor and Hevner, 2013; Reubens, 2016), therefore it can be generalised in another research. We follow similar steps of DSR (Horváth, 2007; Hevner and Chatterjee, 2010), as follows:

**Step 1**: Identify the problem and the importance of the solution. The study problem and its importance is identified in the introduction section. Based on this, the related literature is presented and the main hypothesis of the study is structured. Can disclosure tone be a source of prediction of banks' financial performance? This question is the main concern of the current study. Therefore, how can we prepare predictive models? What are the variables that should be included in these models and what about the validation of these models?

**Step 2**: Design predictive models. This is conducted through three stages. In stage 1, independent variables of the models are identified (financial and non-financial variables) and data are collected. In stage 2, the dependent variable of predictive models, EPS, is selected. Stage 3 involves selecting the appropriate statistical technique to build predictive models, such as machine learning methods. Then, statistical tests on the validity of variables in the proposed models are conducted.

**Step 3**: Develop the models. In stage 4, the current study has developed predictive models using three different methods of machine learning, namely LDA, QDA, and RF.

**Step 4**: Evaluate validation of the predictive models and generalise to other applications. In stage 5, different statistical tests should be used to identify the validity of the predictive models. Consequently, these models can be used in other applications. Details on these steps and their stages are presented in the following paragraphs.

A practical framework consisting of five stages is designed (the first three stages are included in step 2; stage 4 is employed in step 3, while stage 5 is conducted in step 4), as shown in Figure 1.

[Figure 1 about here]

Figure 1 shows that the research framework consists of five stages: collecting features, EPS classification (EPSC), variable selection and model preparation, building training model, predicting, and evaluation process.

### *Stage 1: Collecting features and variables of the study*

### *Data collection*

We select the banking sector in emerging markets since banks have the highest market capitalisation in stock exchanges of these markets. Consequently, banks' annual reports were collected for a sample of 63 listed conventional banks covering 10 years (2008-2017) from eight emerging countries: Bahrain, Kuwait, United Arab Emirates, Saudi Arabia, Qatar, Jordan, and Egypt. The reason for choosing these eight countries is due to the similarity and homogeneity between these countries in terms of economic conditions, culture, and language, as well as customs and traditions. The final sample consists of a total of 630 annual bank reports. Islamic banks are excluded because of their different nature. The selection of banks is based on the availability of data during the study period. Information on the quantitative properties of banks has been collected from different sources such as the eight websites of emerging markets, websites of listed banks, and their financial statements. Table 1 shows the distribution of the sample by country and year.

[Table 1 about here]

### *Variables of the study*

The current study uses two groups of variables to build its predictive models, as follows:

### *First group: Disclosure tone in banks' annual reports (as non-financial variables)*

We use Loughran and McDonald's (LM) (2011) dictionary to analyse disclosure tone in banks' annual reports for several reasons. First, the LM dictionary has been used by several researchers (Feldman et al., 2010; Li, 2010; Garcia, 2013; Liu and McConnell, 2013; Huang et al., 2014; Baginski et al., 2018) in accounting and finance because it is a comprehensive dictionary related to business areas, unlike others such as the Henry (2008) list and Harvard IV-4, which are general lists. Kearney and Liu (2014) point out that "the LM lists have become predominant in more recent studies" (p.175). Second, the LM dictionary is based on the analysis of textual contents of a large sample of two forms (10-K and 10-Q) from 1994-2008 in an attempt to identify the language that managers use in their communication, therefore it deals with the actual language of managers. In addition, it has an extensive list of words; for example, it includes 354 positive and 2,329 negative words. We employ the sentiment analysis of disclosure tone in banks' annual reports using the LM dictionary. Six tone variables are identified to include in our models, namely 'positive', 'negative', 'litigious', 'uncertain', 'constraining', and 'superfluous'. These can be considered as the bag-of-words, and the model decides the importance of individual words for performance prediction. The current study expects that disclosure tones of banks' annual reports (as non-financial information) are forecastable, where disclosure tone in period $t$ can be used in the prediction of bank performance in the period $t + 1$.

***Second group: The quantitative properties of banks (as financial variables)***

Our model requires accounting and financial market variables, therefore we consider prior studies that examine the association between firm performance and corporate disclosure as well as firm attributes. For example, Davis and Tama-Sweet (2012) investigate managers' use of language through earnings press releases and MD&A disclosures with different variables such as firm performance, firm size, leverage, firm beta, loss, the market value of equity, cash flow from operations, and other variables. Keusch et al. (2012) use several control variables in their study on managers' incentives for selecting voluntary disclosures such as firm size, financial leverage, profitability, and change in performance.

Moreover, several studies use different financial ratios in building models to predict many purposes such as corporate failure, bankruptcy, financial disasters, and the financial performance of the firm. For example, Appiah and Abor (2009) use several financial ratios such as liquidity, leverage, and profitability ratios to build their model. In Jordan, Al-Khatib and Al-Horani (2012) use a set of 24 financial ratios to predict the financial distress of a sample of listed companies including leverage and liquidity ratios. In the UK, Smith and Taffler (2000) use Z-score to predict firm failure. Additionally, Kloptchenko et al. (2002) use seven ratios to predict corporate financial performance. In the USA, Balakrishnan et al. (2010) use firm size, market-to-book ratio, and other financial variables in their predictive model. Many prior studies (Altman et al., 1994, 2010; Guo et al., 2006) provide evidence that firm size plays a vital role in making several decisions in the firm and can impact firm profitability. Dias and Matias-Fonseca (2010) use 31 financial ratios to predict corporate performance including leverage and others, similar to Onder and Altintas (2017). Finally, based on the objective of the current study and previous arguments, we consider five financial variables that reflect the quantitative properties of banks: bank size, financial leverage, market-to-book ratio, bank age, and company risk measured by beta. Table 2 shows a summary of the variables used in the study.

[Table 2 about here]

### *Stage 2: EPS classification (EPSC)*

In accounting literature, several measures of financial performance are used such as ROA, ROE, EPS, net income, stock returns, and others (Kloptchenko et al., 2002; Zhang et al., 2004). EPS is the portion of a firm profit allocated to each share of common stocks. It is an indicator of a firm's profitability that can be used in comparison to performance among different firms for the same period. EPS is used by several studies to predict corporate financial performance. For example, Qiu (2007) provides evidence on EPS that is significantly and consistently better than other measures of financial performance, such as ROE and stock return measures, in his proposed models of prediction. Using EPS as a measure of financial performance, Zhang et al. (2004) structure a variety of neural network models. The current study selects EPS as an indicator of the financial performance of a bank. We follow the classification of EPS that is conducted by Qiu (2007), Balakrishnan et al. (2010), and Qiu et al. (2014) who categorise the data of each firm in a year corresponding into three categories (based on 25-50-25%). Consequently, the banks that have EPS over $q_{0.75}$ (where q-quantiles refer to the values that divide a set of data into q subsets, for example, $q_{0.25}$ contains all the values in the lowest quarter of a data) are considered top performance banks (Top); banks that have EPS between $q_{0.25}$ and $q_{0.75}$ are considered middle-performance banks (Mid); and finally, banks that have EPS less than $q_{0.25}$ are considered bottom performance banks (Bot). Based on EPS classification (EPSC), a bank that lies in any of the three categories (Top, Mid, Bot) for a selected year depends on its performance compared with other banks in all years of the study. The current study uses a sample of 63 conventional banks with total observations of 630 bank-years. Because of the low number of listed banks in these markets and the absence of data for more than 10 years for these banks, we decide to build predictive models that predict banks' financial performance in year $t + 1$ (year 2017) using the previous years $t$ (in our case, years from 2008 to 2016). In other words, a training model is built that is based on the data from nine years (from 2008 to 2016) then, we predict classification for bank performance in the year 2017. This decision is taken to improve the quality of prediction in our models. Some studies that have a large sample of firms (e.g., Qiu, 2007; Balakrishnan et al., 2010) use year $t + 1$ to predict a company's change in performance the following year based on a previous year $t$.

***Stage 3: Feature selection (model preparation)***

Feature selection is an important process in machine learning methods. The increase of the accuracy of predictive models depends on the better selection of the independent variables (Hastie et al., 2009; Genuer, 2010). We employ two tests to check the validity of variables in the proposed models, which are the correlation analysis and Boruta algorithm.

Table 3 shows the correlation among the independent variables of the study. The highest correlation is 0.93 between NEGT and LITT variables and this may cause multicollinearity among the independent variables. However, a high correlation among the independent variables may affect the accuracy of the proposed models (James et al., 2013; Tharwat, 2016).

[Table 3 about here]

To detect the multicollinearity problem, the current study computes the variance inflation factor (VIF). If the VIF value exceeds 10, this indicates multicollinearity problems (James et al., 2013). Table 4 shows the VIF for all independent variables of the study. It should be noted that the VIF value exceeds 10 for LITT (10.49) and NEGT variables (13.57).

After deleting the NEGT variable, there are no further multicollinearity problems as all values are less than 10. After deleting the LITT variable, the multicollinearity problem still exists with a value of more than 10 for the NEGT variable. This indicates that NEGT and LITT variables are reflecting the same information. Therefore, the NEGT variable is deleted in subsequent analysis.

[Table 4 about here]

To achieve a better selection of variables in predictive models in our study, we use the Boruta algorithm to identify and test important variables that are statistically significant (Kursa et al., 2010). Table 5 shows the results.

[Table 5 about here]

It can be noted that all the independent variables are confirmed to be selected in our models because they are significant at the level of 0.01. In other words, all 10 independent variables help to predict the dependent variable.

***Stage 4: Building a training model then testing the model (predictive model)***

A training model is built using the data of the previous nine years $t$ (2008 to 2016) to predict the EPSC of bank performance in year $t+1$ (year 2017). The current study develops models to predict EPSC for a sample of 63 listed conventional banks, as shown in Table 1. Our main model consists of a group of five disclosure tones after deleting the NEGT variable, and five quantitative properties of banks using three different methods of machine learning, namely LDA, QDA, and RF.

The predictive model of financial performance (using LDA, QDA, and RF) is as follows:

$$\text{EPSC} = f(\text{Financial Variables} + \text{non-Financial Variables})$$

Where EPSC, the classification of EPS, is used as an indicator of bank financial performance.

$$\text{Financial Variables} = (LogBSIZ, LEVR, BETA, MKBK, BAGE)$$

and

$$\text{non-Financial Variables} = (LITT, POST, UNCT, CONT, SUP)$$

To test the main hypothesis of the study – whether the quality of predictive models of financial performance can be improved by incorporating disclosure tone variables – we run the model with financial variables only using the same three methods (LDA, QDA, and RF), as follows:

$$\text{EPSC} \ = \ f(LogBSIZ, LEVR, BETA, MKBK, BAGE)$$

***Stage 5: Evaluation process of the predictive models in the study***

Most machine learning techniques use mean square error and confusion matrix methods to compare models and evaluate them. In the classification problem, the confusion matrix plays an important role in evaluating the model (James et al., 2013). Accuracy is one of the most important characteristics used to measure the quality of models. However, in the case of unequal numbers in classes or in cases with more than two classes in the data, it can be a misleading measure. The confusion matrix can give a better understanding of the performance of the classification model because it uses different characteristics to measure the quality of models (Altman and Bland, 1994a, 1994b; Kuhn, 2008).

These measures are accuracy, Kappa coefficient, sensitivity, specificity, positive predicted value, negative predicted value, harmonic mean, and balance accuracy, where *accuracy* reflects the percentage of acceptance correct classification and prediction. *Kappa coefficient* reflects how often the class performs if it compares with its performance by chance. These two measures are used for the entire model. The measures that are used for the classes are *sensitivity* (recall), that reflects the percentage of acceptance correct classification for a given observed class; *specificity*, that reflects the percentage of rejection incorrect classification for a given observed class; *positive predicted value* (precision-PPV), that reflects the percentage of acceptance correct prediction for a given predicted class; *negative predicted value* (NPV), that reflects the percentage of rejection incorrect prediction for a given predicted class; the *harmonic mean* ($F_1$), that reflects the balance between precision and recall; and *balance accuracy*, that reflects the mean of sensitivity and specificity (see Altman and Bland, 1994a; 1994b; Donner and Klar, 1996). The classification model predicts data points as top-performing, medium-performing, and below-performing according to comparison between the prediction and true labels to determine the model accuracy, and other measures such as sensitivity and specificity. The current study uses the above measures to evaluate the performance of its proposed models to predict the financial performance of the bank.

Finally, the proposed framework in our study provides many research opportunities that may be tested in the future by other researchers. The applied framework presents new horizons and trends in the field of forecasting. For example, future studies could use our study’s framework to predict stock market prices and the efficiency of investment projects. Moreover, the results of our study may be of interest to many parties. For example, shareholders, investors, creditors, and others can benefit from these results. They can use the framework from our study in forecasting corporate financial performance, which helps to improve the quality of investment decisions. In addition, our findings can offer several practical implications. For instance, this study provides evidence on disclosure tone as qualitative variables increase the quality of predictive models for corporate financial performance. Consequently, it suggests that the analysis of disclosure tone can be used as a complementary or diagnostic approach rather than an alternative by different parties such as analysts, investors, auditors, and others in making their decisions.

# 5. Empirical analysis

## 5.1 Descriptive analysis

Table 6 presents the descriptive statistics. It shows that the mean and median for LogBSIZ are very close with a small coefficient of variation (Coef.Var) (0.20) among the data. A similar pattern can be noted for variables such as LEVR, BETA, and BAGE. In contrast, the mean and median of disclosure tone variables (LITT, POST, UNCT, CONT, and SUPF) are not close and have a high variation. This indicates that the distribution of these variables is mostly non-symmetric. In the current study, all the analyses are done through R-Software using the CARET package (Kuhn, 2008).

[Table 6 about here]

Figure 2 shows the distribution of the independent variables across EPSC using Kernel density estimates. The plot suggests three different distributions for LogFSIZ with bimodality in Mid class. This indicates that the LogBSIZ has different means in each group and consequently has a great effect on EPSC. Moreover, there are different shapes for the variables BETA, UNCT, and CONT within the groups. SUPF, POST, and LITT are the variables that have the most similar shapes in the three groups.

[Figure 2 about here]

## 5.2 Building predictive models with financial variables only (LDA, QDA, and RF)

The current study seeks to examine whether the disclosure tone contained in the banks' annual reports can be a source of information about the banks' future performance. Therefore, it is suggested that the quality of predictive models of financial performance can be improved by incorporating disclosure tone in these models. To test the validity of this hypothesis, we run three models (1, 2 and 3) using LDA, QDA, and RF that contained financial variables only that reflect the quantitative properties of banks. Then, the same method is used to build predictive models that include both financial and non-financial variables. Models that contain financial variables only are structured as follows:

$$\text{Performance} = f(\text{Financial Variables})$$

Where

$$\text{Financial Variables} = (LogBSIZ, LEVR, BETA, MKBK, BAGE)$$

The main financial model is

$$\text{EPSC} \ = \ f(LogBSIZ, LEVR, BETA, MKBK, BAGE)$$

### 5.2.1 *LDA method (Model 1: Financial variables only)*

The results of LDA Model 1 are given in Table 7. The overall accuracy of Model 1 is 70% with the confidence interval ranging from 0.58 to 0.81. The p-value of the model is 0, indicating the accuracy is different from the no information rate that reflects the highest proportion of the observed categories. The Kappa value is 0.53 which is quite low. In general, LDA Model 1 does best in the Top class in terms of recall where 95% are in Top (versus 81% Mid and 28% Bot). The model shows 100% specificity for Bot class (versus 62% Mid and 89% Top), 100% precision for Bot class (versus 61% Mid and 78% Top) and 98% rejection of the incorrect prediction for Top class (versus 82% Mid and 78% Bot). Moreover, the Top class is still better than Mid and Bot classes in terms of general measures, namely $F_1$ and balance accuracy measures. The Top class (86%) has a balance between acceptance of the correct prediction and acceptance of the correct classification (versus 70% Mid and 43% Bot) and has a 92% average between acceptance of the correct classification and rejection of the incorrect classification (versus 72% Mid and 64% Bot).

[Table 7 about here]

The important variables of LDA Model 1 are shown in Figure 3. The process of identifying the important variables describes how much the accuracy of the predictive model relies on the information in each feature (Hastie et al., 2009). In LDA Model 1, the most important variable is LogBSIZ followed by BAGE, LEVR, MKBK, and BETA.

### 5.2.2 *QDA method (Model 2: Financial variables only)*

Table 8 provides the results of QDA Model 2 in addition to the confusion matrix, overall and class performance. We run the same combination of variables in Model 1 using the QDA method. The overall accuracy of QDA Model 2 is 72%, with the confidence interval ranging from 0.59 to 0.82. This is almost the same as the LDA method. The p-value of the model is

0, indicating that the accuracy is different from the no information rate. The Kappa value is 0.56, with the p-value for the McNemar test being 0.19, which does not support rejecting equal row and column marginals. In general, QDA Model 2 does best for the Top class in terms of $F_1$ balance between PPV, sensitivity is 79% (versus 72% Mid and 62% Bot), and balance accuracy measure is 85% (versus 75% Mid and 73% Bot). Other measures include recall at 79% (versus 81% Mid and 50% Bot), specificity at 91% (versus 68% Mid and 96% Bot), precision at 79% (versus 65% Mid and 82% Bot), and NPV at 91% (versus 83% Mid and 83% Bot).

[Table 8 about here]

The important variables plot for the QDA method is shown in Figure 3. The most important variable is LogBSIZ, followed by BAGE, LEVR, MKBK, and BETA.  Surprisingly, the rank of the important variables is similar to the LDA method.

### 5.2.3 *RF method (Model 3: Financial variables only)*

The results of RF Model 3 are given in Table 9. The accuracy of RF Model 3 is 81%, Kappa is 71%. Moreover, the mean $F_1$ for RF Model 3 is 81%, the mean balance accuracy is 85%, the mean PPV is 83%, and the mean NPV is 90%. The important variables plot for the RF method is shown in Figure 3. It can be noted that the most important variable is LogBSIZ, followed by BAGE, LEVR, BETA, and MKBK. The order of variables in the three models is similar, except for the order of BETA and MKBK in the RF method.

[Table 9 about here]

The results of LDA, QDA, and RF methods are summarised in Table 10. It can be noted that the results of RF outperform those of LDA and QDA in terms of accuracy, Kappa, mean $\boldsymbol{F_1}$, mean balance accuracy, mean PPV, and mean NPV. Figure 3 shows the important variables using LDA, QDA, and RF methods.

[Table 10 about here]

[Figure 3 about here]

## 5.3 Building predictive models containing financial and non-financial variables

### *5.3.1 LDA Model 4 (Financial and tone variables)*

The results of LDA Model 4 are given in Table 11. The overall accuracy of LDA Model 4 is 70%, with the confidence interval ranging from 0.58 to 0.81. The p-value of the model is 0, indicating the accuracy is different from the no information rate. The Kappa value is 0.54 which is quite low, with the p-value for the McNemar test being 0.16, failing to reject equal proportions of classifiers. In general, LDA Model 4 does best in the Top class, with 89% acceptance of the correct classification (versus 78% Mid and 39% Bot), 89% rejection of the incorrect classification (versus 68% Mid and 96% Bot), 77% acceptance of the correct prediction (versus 64% Mid and 78% Bot) and 95% rejection of the incorrect prediction (versus 81% Mid and 80% Bot). Moreover, the Top class is still better than Mid and Bot classes in terms of general measures, namely $F_1$ and balance accuracy measures. The Top class (83%) has a balance between acceptance of the correct prediction and acceptance of the correct classification (versus 70% Mid and 52% Top) and it has an 89% average between acceptance of the correct classification and rejection of the incorrect classification (versus 73% Mid and 67% Bot).

[Table 11 about here]

The important variables of LDA Model 4 are shown in Figure 4. In LDA Model 4, the most important variable is LogBSIZ, followed by BAGE and UNCT. Surprisingly, the POST variable is the least important. It can be noted that UNCT, the disclosure tone variable, comes before financial variables such as LEVER, MKBK, and BETA, indicating the importance of non-financial variables in predicting performance.

### 5.3.2 *QDA Model 5* (*Financial and tone variables*)

The results of QDA Model 5 are given in Table 12. We run the same combination of variables as in Model 4 using the QDA method. The overall accuracy of Model 2 is 72%, with the confidence interval ranging from 0.59 to 0.82. This is almost the same as the LDA method. The p-value of the model is 0, indicating that the accuracy is different from the no information rate. The Kappa value is 0.56, with a p-value for the McNemar test of 0.19 – this does not support rejecting equal row and column marginals. In general, QDA Model 5 does best in the Top class in terms of $F_1$, balance between PPV and sensitivity is 79% (versus 72% Mid and 62% Bot), and balance accuracy measure is 85% (versus 75% Mid and 73% Bot). Other measures include recall at 79% (versus 81% Mid and 50% Bot), specificity at

91% (versus 68% Mid and 96% Bot), precision at 79% (versus 65% Mid and 82% Bot), and NPV at 91% (versus 83% Mid and 83% Bot).

[Table 12 about here]

The important variables plot for the QDA method is shown in Figure 4. The most important variable is LogBSIZ, followed by BAGE and UNCT. Surprisingly, the rank of the important variables is the same as the LDA method.

### 5.3.3 *RF Model 6* (*Financial and tone variables*)

The results of RF Model 6 are given in Table 13. Comparing the results of RF with the results of other methods such as LDA and QDA, RF results are much better. The overall accuracy of Model 6 is 86%, with the confidence interval ranging from 0.75 to 0.93. The p-value of the model is 0, indicating that the accuracy is different from the no information rate. The Kappa value is 0.78 – much higher than LDA and QDA methods. Moreover, RF Model 6 does well in all classes in terms of $F_1$ (88%, 85%, and 86%, respectively) and balance accuracy (91%, 87%, and 88%, respectively). In addition, the acceptance correct classification is 93% for Mid class (versus 83% Bot and 79% Top), rejection for the incorrect classification is 98% for Bot and Top classes (versus 81% Mid), acceptance for the correct prediction is 94% for Bot and Top classes (versus 78% Mid) and lastly, the rejection for the incorrect prediction is 94% for Bot and Mid classes and 92% for Top class.

[Table 13 about here]

The important variables plot for RF Model 6 is shown in Figure 4. The most important variable is LogBSIZ, followed by BAGE, LEVER, and UNCT. The variable SUP is not important at all.

The UNCT variable ranks fourth before BETA, MKBK, and other non-financial variables. Surprisingly, the variable POST is ranked penultimately. This may indicate that stakeholders need more information about uncertainty than positive news. This may be explained by a general fear of the future due to the uncertainty of what may happen. Risks, losses, and other problems may occur in the future, therefore if an individual can obtain information to explain this uncertainty they are more likely to feel comfortable about it. The absence of good news about the future is something that may not bother investors and shareholders as much as the absence of bad news or uncertainty.

[Figure 4 about here]

It is worth mentioning that RF has one tuning parameter that assists in choosing the optimal model in terms of accuracy and Kappa (randomly selected predictors; see Kuhn, 2008), while there is no tuning parameter for LDA or QDA. It can be noted that the value 4 gives the best level of accuracy for the RF method, as shown in Figure 5.

[Figure 5 about here]

Table 14 summarises the results of LDA, QDA, and RF methods. It is clear that the results of RF outperform the results of both LDA and QDA in terms of accuracy, Kappa, mean $F_1$, mean balance accuracy, mean PPV, and mean NPV.

[Table 14 about here]

## 6. Comparison among models

In terms of the best RF model results, comparison between the results of RF Model 6 (as the best model that contains both disclosure tone variables and financial variables) and RF Model 3 (that contains only the financial variable) shows that the accuracy is 81% for RF Model 3 versus 86% for RF Model 6, and Kappa is 71% versus 78%, respectively. Moreover, the mean $F_1$ for RF Model 3 is 81% versus 86% in RF Model 6, mean balance accuracy is 85% versus 89%, mean PPV is 83% versus 89%, and the mean NPV is 90% versus 93%. The findings indicate that the existence of non-financial variables in predictive models can improve the quality of these models and increase their performance. Since RF Model 3 and RF Model 6 are used for the same training data, this suggests making inferences on the differences between the two models based on results of 100 resampling. Therefore, $a$ $t$-test is used to evaluate whether there are differences between the two models (Hothorn et al., 2005; Kuhn, 2008). The results of the $t$-test show that the accuracy is 5.974 with a p-value of 0, and Kappa is 5.950 with a p-value of 0. The decision rejects no differences between the two models in terms of the accuracy and Kappa measures. This provides statistical evidence that disclosure tone variables increase predictive models' performance in terms of the accuracy and Kappa measures. Moreover, Figure 6 shows a 97.5% confidence interval for differences between the two models in terms of the accuracy and Kappa measures. Since the intervals lie completely on the positive side, this gives further evidence that the accuracy of predictive models can be increased by incorporating disclosure tone variables as non-financial variables with financial variables.

[Figure 6 about here]

According to the results of RF Model 3 and RF Model 6 (shown in Tables 9 and 13 and Figure 6), H1 in the current study is accepted. This result is consistent with Dias and Matias-Fonseca (2010) who provide evidence on using both quantitative (financial ratios) and qualitative data (positive and negative terms in annual reports) to provide a better indication of the future financial performance of the firm. Comparing our results with prior studies shows interesting observations. First, comparing with de Graaff (2017), who developed the Fuzzy Fingerprints model to predict corporate financial performance using ROE, we note that his model has an accuracy of 0.8333 and a Kappa statistic of 0.4973, while our results for RF Model 6 show that the accuracy is 86% for the year 2017 with a Kappa statistic of 0.78. Second, the results of Qiu et al. (2014), who use the SVM method to build a predictive model of US firms, show that the accuracy of their model is 75% for the year 2002.

## 7. Conclusion

The current study examined whether the textual contents of annual reports such as disclosure tone contain valuable information that can improve the quality and accuracy of predictive models of firm performance. We developed four predictive models of banks' financial performance using three techniques of machine learning: LDA, QDA, and RF. Three of these models were based on a set of financial and non-financial variables, while the fourth model contained only financial variables. We ran these models using a sample of 63 conventional banks from eight emerging markets (Egypt, Jordan, Bahrain, United Arab Emirates, Saudi Arabia, Kuwait, Qatar, and Oman) from 2008 to 2017. The findings of the study reveal that the RF method provides the best predictive model for the variables of our study in terms of the overall accuracy (86%) and Kappa (78%) measures in RF Model 6. LDA Model 1 showed an accuracy of 70% and Kappa of 53%; in addition, QDA Model 2 had an accuracy of 72% and Kappa of 56%. The results of RF Model 3 showed an accuracy of 81% and Kappa of 71%. Moreover, we provided evidence that the incorporation of disclosure tone variables into predictive models with financial variables increased the accuracy and quality of these models. Concerning disclosure tone variables, uncertainty information is the most important variable among the proposed models, indicating that fear of the future is associated with uncertainty status and consequently, it has priority over other tone variables. The most important variable related to the quantitative properties of banks is the size of the bank.

The findings of our study offer practical implications. For instance, the study provides evidence that disclosure tone, as qualitative variables, increases the quality of predictive models for corporate financial performance. Consequently, it suggests that the analysis of disclosure tone can be used as a complementary or diagnostic approach rather than an alternative by different parties such as analysts, investors, auditors, and others in making their decisions.

Of course, this study is not without some limitations. For example, the size of the sample is relatively small, with 63 conventional banks. Future research can use these limitations as a topic for further studies, where sample size can be increased, and different types of business can be examined such as industrial and service companies. In addition, the study combines a group of non-financial variables with financial variables to build predictive models. Future studies could test other combinations of variables such as liquidity, foreign listing, and board characteristics.

**Conflict of Interest:** The authors do not have any conflicts of interest to declare.

**Acknowledgement**

The authors are grateful to the Editor of the International Journal of Disclosure and Governance for their help and support. The authors also would like to thank the anonymous referees for their constructive comments. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## References

Al-khatib, H, B, and Al-Horani, A. (2012) ‘Predicting financial distress of public companies listed in Amman stock exchange’. *European Scientific Journal*, 8(15), pp1-18.

Aly, D., El-Halaby, S., & Hussainey, K. (2018) ‘Tone disclosure and financial performance: Evidence from Egypt’. *Accounting Research Journal*, 31(1), pp.63-74.

Altman, D.G., & Bland, J.M. (1994a) ‘Diagnostic tests. 1: Sensitivity and specificity’. *British Medical Journal*, 308, p.1552.

Altman, D.G., & Bland, J.M. (1994b) ‘Statistics notes: Diagnostic tests 2: Predictive values’. *British Medical Journal*, 309, p.102.

Altman, E.I., Sabato, G., & Wilson, N. (2010) ‘The value of non-financial information in small and medium-sized enterprise risk management’. *Journal of Credit Risk*, 2, pp.95-127.

Altman, E.I., Iwanicz-Drozdowska, M., Laitinen, E.K., & Suvas, A. (2017) ‘Financial distress prediction in an international context: A review and empirical analysis of Altman’s z-score model’. *Journal of International Financial Management & Accounting*, 28, pp.131-171.

Appiah, K.O., & Abor, J. (2009) ‘Predicting corporate failure: Some empirical evidence from the UK’. *Benchmarking: An International Journal*, 16, pp.432-444.

Appiah, K.O., Chizema, A., & Joseph Arthur, J. (2015) ‘Predicting corporate failure: A systematic literature review of methodological issues’. *International Journal of Law and Management,* 57, pp.461-485.

Arslan-Ayaydin, Ö., Boudt, K., & Thewissen, J. (2016) ‘Managers set the tone: Equity incentives and the tone of earnings press releases’. *Journal of Banking & Finance*, 72, pp.S132-S147.

Aziz, M.A., & Dar, H.A. (2006) ‘Predicting corporate bankruptcy: Where do we stand?’. *Corporate Governance*, 6, pp.18-33.

Association for Investment Management and Research (AIMR). (2000) *Corporate Disclosure Survey: A Report to AIMR*. Fleishman-Hillard Research.

Baginski, S.P., Demers, E., Kausar, A., & Yu, Y.J. (2018) ‘Linguistic tone and the small trader’. *Accounting, Organizations and Society*, 68-69, pp.21-37.

Balcaen, S., & Ooghe, H. (2006) ‘35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems’. *The British Accounting Review*, 38, pp.63-93.

Beattie, V.A., & Jones, M.J. (2000) ‘Changing graph use in corporate annual reports: A time-series analysis’. *Contemporary Accounting Research*, 17(2), pp.213-226.

Breiman, L. (2001a) ‘Random forests’. *Machine Learning*, 45, pp.5-32.

Breiman, L. (2001b) ‘Statistical modeling: The two cultures’. *Statistical Science*, 16, pp.199-215.

Breiman, L. (2017) *Classification and regression trees*. Routledge.

Balakrishnan, R., Qiu, X.Y., & Srinivasan, P. (2010) ‘On the predictive ability of narrative disclosures in annual reports’. *European Journal of Operational Research*, 202, pp.789-801.

Campbell, J.L., Chen, H., Dhaliwal, D.S., Lu, H.M., & Steele, L.B. (2014) ‘The information content of mandatory risk factor disclosures in corporate filings’. *Review of Accounting Studies*, 19, pp.396-455.

Connelly, B.L., Certo, S.T., Ireland, R.D., & Reutzel, C.R. (2011) ‘Signaling theory: A review and assessment’. *Journal of Management*, 37(1), pp.39-67.

Chen, K., Chen, T., & Yen, J. (2009) ‘Predicting future earnings change using numeric and textual information in financial reports’. *Proceedings of Intelligence and Security Informatics, Pacific Asia Workshop*, PAISI 2009, Bangkok, Thailand, pp.54-63.

Chou, C., Chang, C.J, Chin, C., & Chiang, W. (2018) ‘Measuring the consistency of quantitative and qualitative information in financial reports: A design science approach’. *Journal of Emerging Technologies in Accounting*, 15, pp.93-109.

Clatworthy, M., & Jones, M. (2003) ‘Financial reporting of good news and bad news: Evidence from accounting narratives’. *Accounting and Business Research*, 33(3), pp.171-185.

Clatworthy, M., & Jones, M. (2006) ‘Differential patterns of textual characteristics and company performance in the chairman’s statement’. *Accounting, Auditing & Accountability Journal*, 19(4), pp.493-511.

Cooper, W., Seiford, L., & Tone, K. (2000) *Data envelopment analysis: A comprehensive text with models, applications, references and DEA solver software*. Boston/Dordrecht/London: Kluwer Academic Publishers.

Davis, A.K., & Tama-Sweet, I. (2012) ‘Managers’ use of language across alternative disclosure outlets: Earnings press releases versus MD&A’. *Contemporary Accounting Research*, 29(3), pp.804-837.

Davis, A., Piger, J., & Sedor, L. (2006) *Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases*. Working paper, Washington University at St. Louis, Federal Reserve Bank of St. Louis and the University of Notre Dame.

de Graaff, R. (2017) *Sentiment analysis of annual reports as a financial performance indicator*. Master of Science in Business Information Systems, Eindhoven University of Technology, Holland.

Dias, W., & Matias-Fonseca, R. (2010) ‘The language of annual reports as an indicator of the organizations’ financial situation’. *International Review of Business Research Papers*, 6, pp.206-215.

Donner, A., & Klar, N. (1996) ‘The statistical analysis of kappa statistics in multiple samples’. *Journal of Clinical Epidemiology*, 49, pp.1053-1058.

Duda, R.O., Hart, P.E., & Stork, D.G. (2012) *Pattern classification*. John Wiley & Sons.

Elshandidy, T., & Shrives, P.J. (2016) ‘Environmental incentives for and usefulness of textual risk reporting: Evidence from Germany’. *The International Journal of Accounting,* 51, pp.464-486.

Falschlunger, L.M., Eisl, C., Losbichler, H., & Greil, A.M. (2015) ‘Impression management in annual reports of the largest European companies: A longitudinal study on graphical representations’. *Journal of Applied Accounting Research*, 16(3), pp.383-399.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010) ‘Management’s tone change, post earnings announcement drift and accruals’. *Review of Accounting Studies,* 15, pp.915-953.

Garcia, D. (2013) ‘Sentiment during recessions’. *Journal of Finance,* 68, pp.1267-1300.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010) ‘Variable selection using random forests’. *Pattern Recognition Letters,* 31, pp.2225-2236.

Gregor, S., & Hevner, A.R. (2013) ‘Positioning and presenting design science research for maximum impact’. *MIS Quarterly*, pp.337-355.

Hackfort, D., Schinke, R., & Strauss, B. (2019) *Dictionary of sport psychology*. Academic Press.

Hahn, R., & Lülfs, R. (2014) ‘Legitimizing negative aspects in gri-oriented sustainability reporting: A qualitative analysis of corporate disclosure strategies’. *Journal of Business Ethics,* 123(3), pp.401-420.

Hastie, T., Tibshirani, R., & Friedman, J. (2009) *The elements of statistical learning: Data mining, inference, and prediction*. Springer Series in Statistics. New York: Springer.

Heinle, M.S. & Smith, K.C. (2017) ‘A theory of risk disclosure’. *Review of Accounting Studies,* 22, pp.1459-1491.

Henry, E. (2006) ‘Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm’. *Journal of Emerging Technologies in Accounting,* 3, pp.1-19.

Henry, E. (2008) ‘Are investors influenced by how earnings press releases are written?’. *Journal of Business Communication,* 45, pp.363-407.

Hevner, A., & Chatterjee, S. (2010) ‘Design science research in information systems’. In: A. Hevner & S. Chatterjee (eds.), *Design research in information systems* (pp.9-22). Boston, MA: Springer.

Hildebrandt, H.W., & Snyder, R.D. (1981) ‘The Pollyanna hypothesis in business writing: Initial results, suggestions for research’. *Journal of Business Communication,* 18, pp.5-15.

Ho, T.K. (1995) ‘Random decision forests’. *Proceedings of the 3rd International Conference on Document Analysis and Recognition,* Montreal, QC, 14-16 August 1995. pp.278-282.

Ho, T.K (1998) ‘The random subspace method for constructing decision forests’. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20(8), pp.832-844.

Ho, C.T., & Zhu, D.S. (2004) ‘Performance measurement of Taiwan’s commercial banks’. *International Journal of Productivity and Performance Management*, 53(5/6), pp.425-434.

Hope, O., Hu, D., & Lu, H. (2016) *The benefits of specific risk-factor disclosures*. Rotman School of Management Working Paper No. 2457045; Singapore Management University School of Accountancy Research Paper No. 2015-35.

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005) ‘The design and analysis of benchmark experiments’. *Journal of Computational and Graphical Statistics*, 14, pp.675-699.

Huang, X., Teoh, S.H., & Zhang, Y. (2014) ‘Tone management’. *The Accounting Review*, 89(3), pp.1083-1113.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013) *An introduction to statistical learning*. Springer.

Jensen, H., & Meckling, W. (1976) ‘Theory of the firm: Managerial behaviour, agency costs and ownership structure’. *Journal of Financial Economics*, 16, pp.305-360.

Kearney, C., & Liu, S. (2014) ‘Textual sentiment in finance: A survey of methods and models’. *International Review of Financial Analysis*, 33, pp.171-185.

Keusch, T., Bollen, L.H.H., & Hassink, H.F.D. (2012) ‘Self-serving bias in annual report narratives: An empirical analysis of the impact of economic crises’. *European Accounting Review*, 21(3), pp.623-648.

Kloptchenko, A., Eklund, T., Back, B., Karlsson, J., Vanharanta, H., & Visa, A. (2002) ‘Combining data and text mining techniques for analysing financial reports’. In: *Proceedings of Eighth Americas Conference on Information Systems*.

Kravet, T., & Muslu, V. (2013) ‘Textual risk disclosures and investors’ risk perceptions’. *Review of Accounting Studies*, 18, pp.1088-1122.

Kuhn, M. (2008) ‘Building predictive models in R using the caret package’. *Journal of Statistical Software*, 28, pp.1-26.

Kursa, M.B., Jankowski, A., & Rudnicki, W.R. (2010) ‘Boruta: A system for feature selection’. *Fundamenta Informaticae*, 101, pp.271-285.

Kursa, M., & Rudnicki, W. (2010) ‘Feature selection with the Boruta package’. *Journal of Statistical Software*, 36(11), pp.1-13.

Leary, M.R. (2001) ‘Impression management, psychology of’. In: J. wright (ed.), *International Encyclopedia of the social & behavioral sciences* (pp.7245-7248). Elsevier.

Leary, M.R., & Kowalski, R.M. (1990) ‘Impression management: A literature review and two component model’. *Psychological Bulletin*, 107(1), pp.34-47.

Lee, A., Lin, J.T., Kao, R., & Chen, K. (2010) ‘An effective clustering approach to stock market prediction’. *PACIS 2010 Proceedings*, p.54.

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

Lev, B. (1989) ‘On the usefulness of earnings: lessons and directions from two decades of empirical research’. *Journal of Accounting Research*, 27(Supplement), pp.153-192.

Li, F. (2006a) *Annual report readability, current earnings and earnings persistence*. Working paper, University of Michigan, Ann Arbor.

Li, F. (2006b) *Do stock market investors understand the risk sentiments of corporate annual reports?* Working paper, University of Michigan, Ann Arbor.

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

Li, F. (2010) ‘The information content of forward-looking statements in corporate filings: A naïve Bayesian machine learning approach’. *Journal of Accounting Research*, 48, pp.1049-1102.

Liu, B., & McConnell, J.J. (2013) ‘The role of the media in corporate governance: Do the media influence managers’ capital allocation decisions?’. *Journal of Financial Economics,*110, pp.1-17.

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

Lu, C., & Chen, T. (2009) ‘A study of applying data mining approach to the information disclosure for Taiwan’s stock market investors’. *Expert Systems with Applications*, 36, pp.3536-3542.

Lukason, O., & Laitinen, E.K. (2019) ‘Firm failure processes and components of failure risk: An analysis of European bankrupt firms’. *Journal of Business Research*, 98, pp.380-390.

Magnusson, C., Arppe, A., Eklund, T., Barbro, B., Vanharanta, H., & Visa, A. (2005) ‘The language of quarterly reports as an indicator of change in the company’s financial status’. *Information & Management*, 42, pp.561-574.

Merkl-Davies, D.M., & Brennan, N.M. (2007) ‘Discretionary disclosure strategies in corporate narratives: Incremental information or impression management?’. *Journal of Accounting Literature*, 27, pp.116-196.

Merkl-Davies, D.M., & Brennan, N.M. (2011) ‘A conceptual framework of impression management: New insights from psychology, sociology and critical perspectives’. *Accounting and Business Research*, 41(5), pp.415-437.

McLachlan, G. (2004) *Discriminant analysis and statistical pattern recognition* (Vol. 544). John Wiley & Sons.

Onder, E., & Altintas, A.T. (2017) ‘Financial performance evaluation of Turkish construction companies in Istanbul Stock Exchange (BIST)’. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, 7, pp.108-113.

Qiu, X.Y. (2007) *On building predictive models with company annual reports* (PhD thesis, University of Iowa, Iowa City, USA).

Qiu, X.Y., Srinivasan, P., & Hu, Y. (2014) ‘Supervised learning models to predict firm performance with annual reports: An empirical study’. *Journal of the American Society for Information Science and Technology,* 65, pp.400-413.

Qiu, X.Y., Srinivasan, P., & Street, N. (2006) ‘Exploring the forecasting potential of company annual reports’. In: *69th Annual Meeting of the American Society for Information Science and Technology (ASIST),* Austin (US), 3-8 November 2006.

Rahman, S. (2012) ‘Impression management motivations, strategies and disclosure credibility of corporate narratives’. *Journal of Management Research*, 4(3).

Ressas, M.S., & Hussainey, K. (2014) ‘Does financial crisis affect financial reporting of good news and bad news?’. *International Journal of Accounting, Auditing and Performance Evaluation*, 10(4), pp.410-429.

Rogers, K., & Grant, J. (1997) ‘Content analysis of information cited in reports of sell-side financial analysts’. *Journal of Financial Statement Analysis*, 3, pp.17-30.

Schleicher, T., & Walker, M. (2010) ‘Bias in the tone of forward-looking narratives’. *Accounting and Business Research,* 40(4), pp.371-390.

Sharma, S. (1995) *Applied multivariate techniques*. John Wiley & Sons, Inc.

Siqueira, L.F., Júnior, R.F.A., de Araújo, A.A., Morais, C.L., & Lima, K.M. (2017) ‘LDA vs. QDA for FT-MIR prostate cancer tissue classification’. *Chemometrics and Intelligent Laboratory Systems*, 162, pp.123-129.

Smith, M., & Taffler, R.J. (2000) ‘The chairman’s statement: A content analysis of discretionary narrative disclosures’. *Accounting Auditing & Accountability Journal*, 13(5), pp.624-647.

Tedeschi, J.T., & Riess, M. (1981) ‘Identities, the phenomenal self, and laboratory research’. In: J.T. Tedeschi (ed.), *Impression management theory and social psychological research* (pp.3-22). New York: Academic Press.

Tharwat, A. (2016) ‘Linear vs. quadratic discriminant analysis classifier: A tutorial’. *International Journal of Applied Pattern Recognition*, 3, pp.145-180.

Yu, Y. Barros, A., Tsai, C., & Liao, K. (2014) ‘A comparison of ratios and data envelopment analysis: Efficiency assessment of Taiwan public listed companies’. *International Journal of Academic Research in Accounting, Finance and Management Sciences,* 4(1), pp.212-219.

Zhang, W. Cao, Q., & Schniederjans, M. (2004) ‘Neural network earnings per share forecasting models: A comparative analysis of alternative methods’. *Decision Sciences*, 5, pp.205-237.

Table 1: Distribution of listed conventional banks by country and year

| | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bahrain | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 40 |
| Egypt | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 90 |
| Emirates | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 13 | 13 | 114 |
| Jordan | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 120 |
| Kuwait | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 80 |
| Oman | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 6 | 6 | 68 |
| Qatar | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 52 |
| Saudi | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 5 | 5 | 66 |
| Total | 63 | 63 | 63 | 63 | 63 | 63 | 63 | 63 | 63 | 63 | 630 |

Table 2: Variables included in the predictive models of financial performance

| Variables | Symbol | Measure |
|---|---|---|
| **Financial performance**<br>(Earnings per share) | EPS | Earnings/total number of the outstanding *shares* of common *stock*. |
| **Quantitative properties of banks (financial variables)** | | |
| Bank size | LogBSIZ | The natural logarithm of total assets. |
| Financial leverage | LEVR | Total liabilities/total assets |
| Market-to-book ratio | MKBK | Firm book value to its market value |
| Beta of the company | BETA | A measure of a stock's volatility in relation to the market. |
| Bank age | BAGE | The number of years from the date of establishment of the company. |
| **Disclosure tone (non-financial variables)** | | |
| Positive tone | POST | The number of positive words in annual reports. |
| Negative tone | NEGT | The number of negative words in annual reports. |
| Constraining | $CONT$ | The number of constraining words in annual reports. |
| Uncertainty | UNCT | The number of uncertainty words in annual reports. |
| Litigious | LITT | The number of litigious words in annual reports. |
| Superfluous | $SUP$ | The number of superfluous words in annual reports. |

Table 3: The correlation matrix among the study's variables

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LogBSIZ | 1.00 | 0.09$^{**}$ | 0.19$^{***}$ | 0.08$^{**}$ | 0.26$^{***}$ | -0.08$^{**}$ | 0.01 | -0.02 | 0.01 | -0.06 | 0.16$^{***}$ |
| **LEVR** | **0.09$^{**}$** | **1.00** | **0.05** | **0.10$^{**}$** | **0.00** | **-0.03** | **0.01** | **0.03** | **0.00** | **-0.06** | **-0.03** |
| BETA | 0.19$^{***}$ | 0.05 | 1.00 | -0.14$^{***}$ | -0.22$^{***}$ | 0.08$^{**}$ | 0.21$^{***}$ | 0.15$^{***}$ | 0.24$^{***}$ | 0.22$^{***}$ | 0.12$^{***}$ |
| MKBK | 0.08$^{**}$ | 0.10$^{**}$ | -0.14$^{***}$ | 1.00 | -0.05 | -0.02 | -0.05 | -0.04 | -0.02 | -0.17$^{***}$ | -0.01 |
| BAGE | 0.26$^{***}$ | 0.00 | -0.22$^{***}$ | -0.05 | 1.00 | 0.05 | 0.00 | 0.07 | -0.03 | -0.15$^{***}$ | -0.05 |
| LITT | -0.08$^{**}$ | -0.03 | 0.08$^{**}$ | -0.02 | 0.05 | 1.00 | 0.93$^{***}$ | 0.87$^{***}$ | 0.88$^{***}$ | 0.14$^{***}$ | 0.03 |
| NEGT | 0.01 | 0.01 | 0.21$^{***}$ | -0.05 | 0.00 | 0.93$^{***}$ | 1.00 | 0.86$^{***}$ | 0.90$^{***}$ | 0.21$^{***}$ | 0.05 |
| POST | -0.02 | 0.03 | 0.15$^{***}$ | -0.04 | 0.07$^{*}$ | 0.87$^{***}$ | 0.86$^{***}$ | 1.00 | 0.78$^{***}$ | 0.12$^{***}$ | 0.00 |
| UNCT | 0.01 | 0.00 | 0.24$^{***}$ | -0.02 | -0.03 | 0.88$^{***}$ | 0.90$^{***}$ | 0.78$^{***}$ | 1.00 | 0.26$^{***}$ | 0.11$^{***}$ |
| CONT | -0.06 | -0.06 | 0.22$^{***}$ | -0.17$^{***}$ | -0.15$^{***}$ | 0.14$^{***}$ | 0.21$^{***}$ | 0.12$^{***}$ | 0.26$^{***}$ | 1.00 | 0.36$^{***}$ |
| SUPF | 0.16$^{***}$ | -0.03 | 0.12$^{***}$ | -0.01 | -0.05 | 0.03 | 0.05 | 0.00 | 0.11$^{***}$ | 0.36$^{***}$ | 1.00 |

Note: *** significant at the level of 0.01, ** significant at the level of 0.05, * significant at the level of 0.10.

Table 4 The VIF for all independent variables of the study

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | VIF for all independent variables | | | | | | | |
| LogBSIZ | LEVR | BETA | MKBK | BAGE | LITT | NEGT | POST | UNCT | CONT | SUPF |
| 1.20 | 1.04 | 1.35 | 1.10 | 1.23 | 10.49 | 13.57 | 4.83 | 7.54 | 7.54 | 1.22 |
| | | | VIF for independent variables after deleting NEGT variable | | | | | | | |
| 1.26 | 1.03 | 1.33 | 1.08 | 1.22 | 8.03 | delete | 4.44 | 5.31 | 1.33 | 1.21 |
| | | | VIF for independent variables after deleting LITT variable | | | | | | | |
| 1.23 | 1.02 | 1.26 | 1.09 | 1.22 | delete | 10.39 | 4.12 | 7.14 | 1.31 | 1.28 |

Table 5: The selection of independent variables using Boruta algorithm

| Variables | Mean Imp | Median Imp | Min Imp | Max Imp | decision |
|---|---|---|---|---|---|
| LogBSIZ | 54.832 | 54.817 | 52.762 | 56.157 | Confirmed |
| LEVR | 26.903 | 27.436 | 23.915 | 28.012 | Confirmed |
| BETA | 24.900 | 25.009 | 22.831 | 26.372 | Confirmed |
| MKBK | 18.327 | 18.097 | 15.854 | 21.221 | Confirmed |
| BAGE | 30.448 | 30.719 | 26.947 | 32.900 | Confirmed |
| LITT | 17.126 | 16.940 | 15.437 | 18.753 | Confirmed |
| POST | 19.069 | 19.073 | 17.765 | 20.290 | Confirmed |
| UNCT | 29.530 | 29.708 | 27.966 | 31.133 | Confirmed |
| CONT | 22.724 | 22.856 | 21.415 | 23.539 | Confirmed |
| SUPF | 12.887 | 12.737 | 12.079 | 13.652 | Confirmed |

(*) Important: Imp

Table 6: Descriptive statistics for the study's variables

| Variables | Median | Mean | SD | Coef.Var |
|---|---|---|---|---|
| LogBSIZ | 3.930 | 4.000 | 0.790 | 0.200 |
| LEVR | 7.620 | 7.910 | 2.370 | 0.300 |
| BETA | 0.740 | 0.720 | 0.370 | 0.520 |
| MKBK | 1.230 | 1.370 | 0.700 | 0.510 |
| BAGE | 36.500 | 35.610 | 13.810 | 0.390 |
| LITT | 101.000 | 115.000 | 128.630 | 1.120 |
| POST | 92.000 | 118.250 | 124.850 | 1.060 |
| UNCT | 292.500 | 351.470 | 333.790 | 0.950 |
| CONT | 140.500 | 147.520 | 93.610 | 0.630 |
| SUPF | 1.000 | 2.110 | 3.890 | 1.840 |

(*) Standard deviation (SD) and coefficient of variation (Coef.Var)

Table 7: The results of LDA Model 1 (the confusion matrix, overall and class performance)

| Confusion matrix | | | | Overall performance | | | Class performance | | |
|---|---|---|---|---|---|---|---|---|---|
| Pred | Bot | Mid | Top | Accuracy | 0.700 | measure | Bot | Mid | Top |
| Bot | 5 | 0 | 0 | 95% CI* | (0.58,0.81) | Sensitivity (recall) | 0.280 | 0.810 | 0.950 |
| Mid | 13 | 22 | 1 | No inf. Rate* | 0.420 | Specificity | 1.000 | 0.620 | 0.890 |
| Top | 0 | 5 | 18 | P-value | 0.000 | PPV* (precision) | 1.000 | 0.610 | 0.780 |
| | | | | Kappa | 0.530 | NPV* | 0.780 | 0.820 | 0.980 |
| | | | | McnemarPvalue | NaN | $F_1$* | 0.430 | 0.700 | 0.860 |
| | | | | | | Balance Accuracy | 0.640 | 0.720 | 0.920 |

*CI: confidence interval. No inf. Rate: no information rate, PPV: positive predicted value, NPV: negative predicted value, F1: harmonic mean

Table 8: The results of QDA Model 2 (the confusion matrix, overall and class performance)

| Confusion matrix | | | | Overall performance | | | Class performance | | |
|---|---|---|---|---|---|---|---|---|---|
| | Bot | Mid | Top | Accuracy | 0.72 | Measure | Bot | Mid | Top |
| Bot | 9 | 2 | 0 | 95% CI* | (0.59,0.82) | Sensitivity (recall) | 0.500 | 0.810 | 0.790 |
| Mid | 8 | 22 | 4 | No inf. Rate* | 0.420 | Specificity | 0.960 | 0.680 | 0.910 |
| Top | 1 | 3 | 15 | P-value | 0.000 | PPV* (precision) | 0.820 | 0.650 | 0.790 |
| | | | | Kappa | 0.560 | NPV* | 0.830 | 0.830 | 0.910 |
| | | | | McnemarPvalue | 0.190 | F1* | 0.620 | 0.720 | 0.790 |
| | | | | | | Balance Accuracy | 0.730 | 0.750 | 0.850 |

*CI: confidence interval. No inf. Rate: no information rate, PPV: positive predicted value, NPV: negative predicted value, F1: harmonic mean

Table 9: The results of RF Model 3 (the confusion matrix, overall and class performance)

| Confusion matrix | | | | Overall performance | | | Class performance | | |
|---|---|---|---|---|---|---|---|---|---|
| | Bot | Mid | Top | Accuracy | 0.810 | Measure | Bot | Mid | Top |
| Bot | 14 | 2 | 0 | 95% CI* | (0.70,0.90) | Sensitivity (recall) | 0.780 | 0.850 | 0.790 |
| Mid | 4 | 23 | 4 | No inf. Rate* | 0.420 | Specificity | 0.960 | 0.780 | 0.960 |
| Top | 0 | 2 | 15 | P-value | 0.000 | PPV* (precision) | 0.880 | 0.740 | 0.880 |
| | | | | Kappa | 0.710 | NPV* | 0.920 | 0.880 | 0.910 |
| | | | | McnemarPvalue | NA | F1* | 0.820 | 0.790 | 0.830 |
| | | | | | | Balance accuracy | 0.870 | 0.820 | 0.870 |

*CI: confidence interval. No inf. Rate: no information rate, PPV: positive predicted value, NPV: negative predicted value, F1: harmonic mean.

Table10: Summary of LDA, QDA and RF Models

| | **Model 1** | **Model 2** | **Model 3** |
|---|---|---|---|
| Measure | LDA | QDA | RF |
| Accuracy | 0.70 | 0.72 | 0.81 |
| Kappa | 0.53 | 0.56 | 0.71 |
| Mean $F_1$ | 0.66 | 0.71 | 0.81 |
| Mean balance accuracy | 0.76 | 0.78 | 0.85 |
| Mean PPV | 0.79 | 0.75 | 0.83 |
| Mean NPV | 0.86 | 0.86 | 0.90 |

Table 11: The results of LDA Model 4 (the confusion matrix, overall and class performance)

| Confusion matrix | | | | Overall performance | | | Class performance | | |
|---|---|---|---|---|---|---|---|---|---|
| pred | Bot | Mid | Top | Accuracy | 0.700 | measure | Bot | Mid | Top |
| Bot | 7 | 2 | 0 | 95% CI* | (0.58,0.81) | Sensitivity (recall) | 0.390 | 0.780 | 0.890 |
| Mid | 10 | 21 | 2 | No inf. Rate* | 0.420 | Specificity | 0.960 | 0.680 | 0.890 |
| Top | 1 | 4 | 17 | P-value | 0.000 | PPV* (precision) | 0.780 | 0.640 | 0.770 |
| | | | | Kappa | 0.540 | NPV* | 0.800 | 0.810 | 0.950 |
| | | | | McnemarPvalue | 0.070 | $F_1$* | 0.520 | 0.700 | 0.830 |
| | | | | | | Balance Accuracy | 0.670 | 0.730 | 0.890 |

*CI: confidence interval. No inf. Rate: no information rate, PPV: positive predicted value, NPV: negative predicted value, F1: harmonic mean

Table 12: The results of QDA Model 5: the confusion matrix, overall and class performance

| Confusion matrix | | | | Overall performance | | | Class performance | | |
|---|---|---|---|---|---|---|---|---|---|
| | Bot | Mid | Top | Accuracy | 0.72 | Measure | Bot | Mid | Top |
| Bot | 9 | 2 | 0 | 95% CI* | (0.59,0.82) | Sensitivity (recall) | 0.500 | 0.810 | 0.790 |
| Mid | 8 | 22 | 4 | No inf. Rate* | 0.420 | Specificity | 0.960 | 0.680 | 0.910 |
| Top | 1 | 3 | 15 | P-value | 0.000 | PPV* (precision) | 0.820 | 0.650 | 0.790 |
| | | | | Kappa | 0.560 | NPV* | 0.830 | 0.830 | 0.910 |
| | | | | McnemarPvalue | 0.190 | F1* | 0.620 | 0.720 | 0.790 |
| | | | | | | Balance Accuracy | 0.730 | 0.750 | 0.850 |

*CI: confidence interval. No inf. Rate: no information rate, PPV: positive predicted value, NPV: negative predicted value, F1: harmonic mean

Table 13: The results of RF Model 6 (the confusion matrix, overall and class performance)

| Confusion matrix | | | | Overall performance | | | Class performance | | |
|---|---|---|---|---|---|---|---|---|---|
| pred | Bot | Mid | Top | Accuracy | 0.860 | Measure | Bot | Mid | Top |
| Bot | 15 | 1 | 0 | 95% CI* | (0.75,0.93) | Sensitivity (recall) | 0.830 | 0.930 | 0.790 |
| Mid | 3 | 25 | 4 | No inf. Rate* | 0.420 | Specificity | 0.980 | 0.810 | 0.980 |
| Top | 0 | 1 | 15 | P-value | 0.000 | PPV* (precision) | 0.940 | 0.780 | 0.940 |
| | | | | Kappa | 0.780 | NPV* | 0.940 | 0.940 | 0.920 |
| | | | | McnemarPvalue | | F1* | 0.880 | 0.850 | 0.860 |
| | | | | | | Balance accuracy | 0.910 | 0.870 | 0.880 |

*CI: confidence interval. No inf. Rate: no information rate, PPV: positive predicted value, NPV: negative predicted value, F1: harmonic mean

Table 14: Summary of LDA, QDA and RF Models

| | **Model 4** | **Model 5** | **Model 6** |
|---|---|---|---|
| Measure | LDA | QDA | RF |
| Accuracy | 0.70 | 0.72 | 0.86 |
| Kappa | 0.54 | 0.56 | 0.78 |
| Mean $F_1$ | 0.68 | 0.71 | 0.86 |
| Mean balance accuracy | 0.76 | 0.78 | 0.89 |
| Mean PPV | 0.73 | 0.75 | 0.89 |
| Mean NPV | 0.85 | 0.86 | 0.93 |

Figure 1: The research framework consists of five stages: collecting features, EPSC, variable selection and model preparation, building training model, and predicting and evaluation process

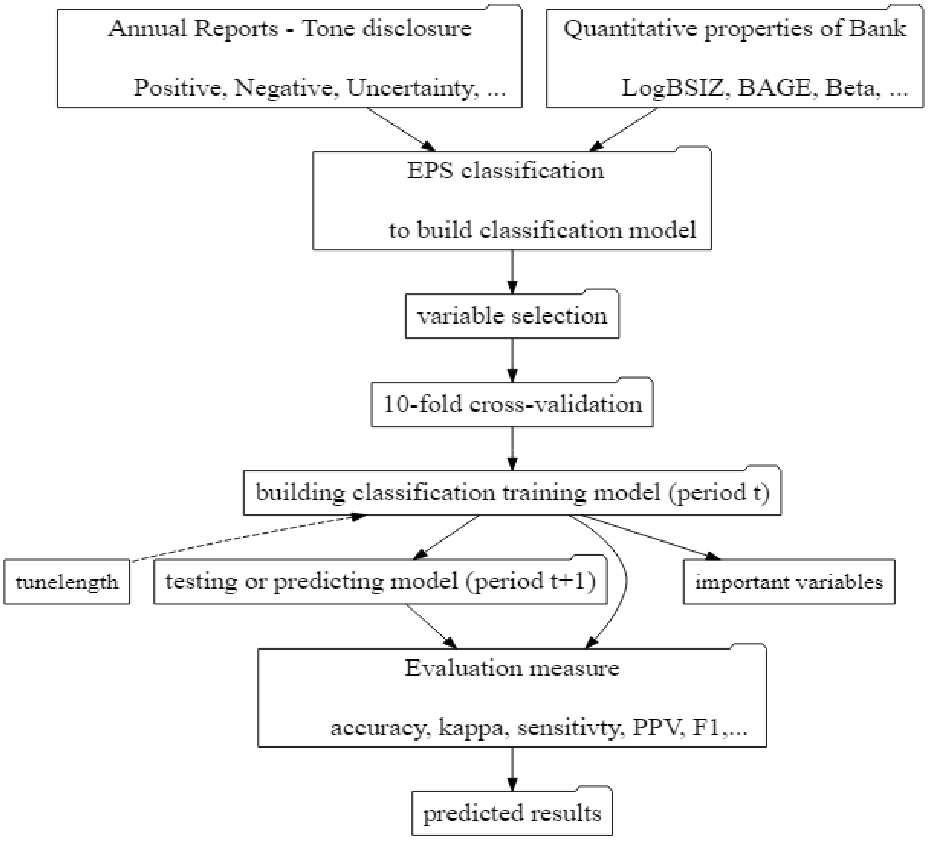

Annual Reports - Tone disclosure
Positive, Negative, Uncertainty, ...
Quantitative properties of Bank
LogBSIZ, BAGE, Beta, ...
EPS classification
to build classification model
variable selection
10-fold cross-validation
building classification training model (period t)
tunelength
testing or predicting model (period t+1)
important variables
Evaluation measure
accuracy, kappa, sensitivty, PPV, F1,...
predicted results

Figure 2:Density plot for independent variables across EPSC

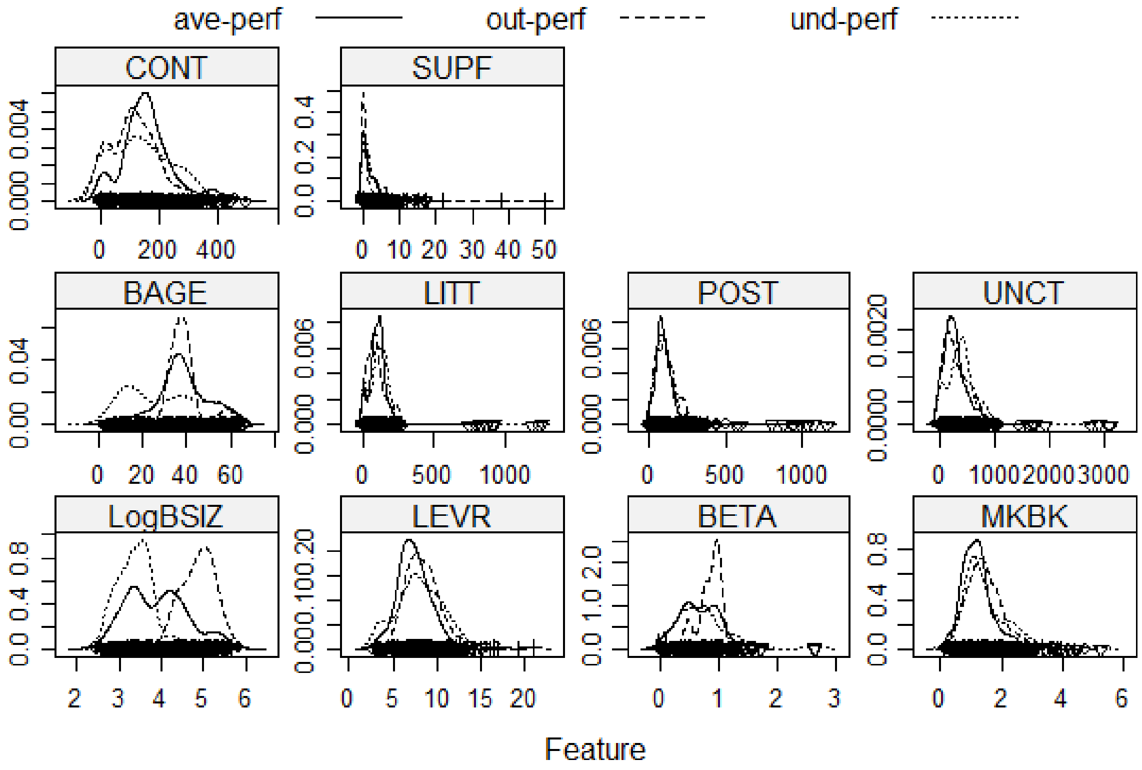


Figure 3: Important variables using LDA, QDA and RF methods

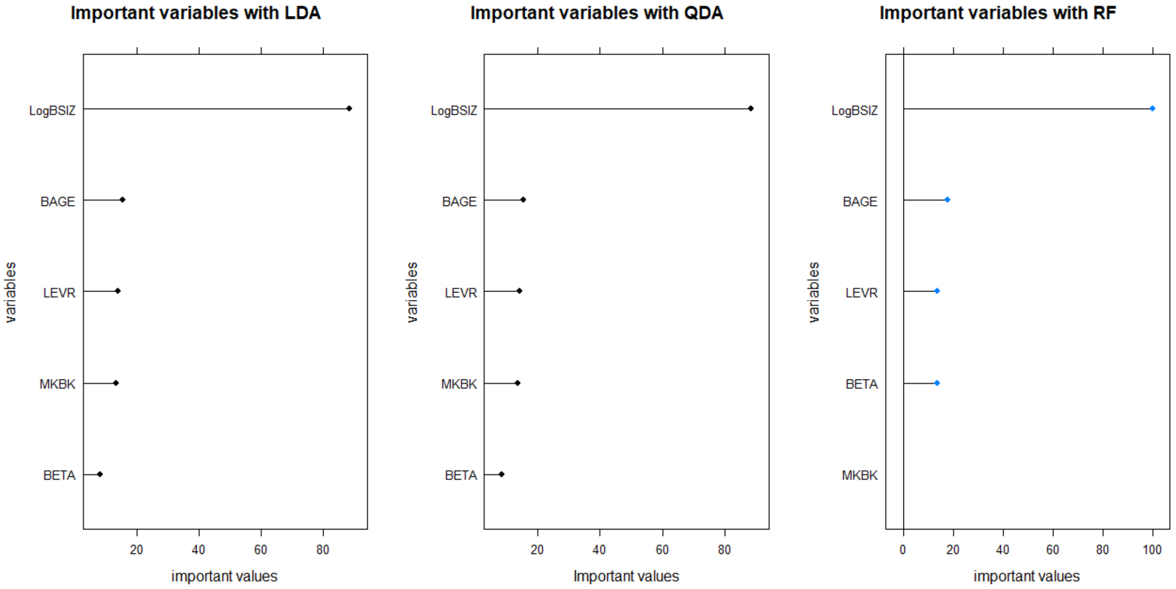


Figure 4: Important variables using LDA, QDA and RF methods

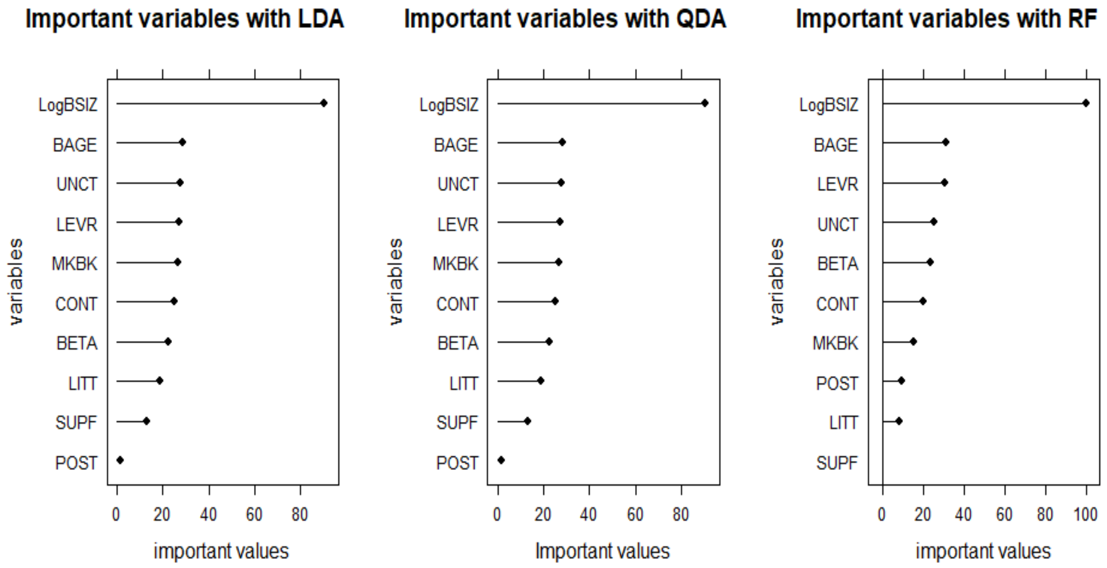


Figure 5: Tuning length for RF method

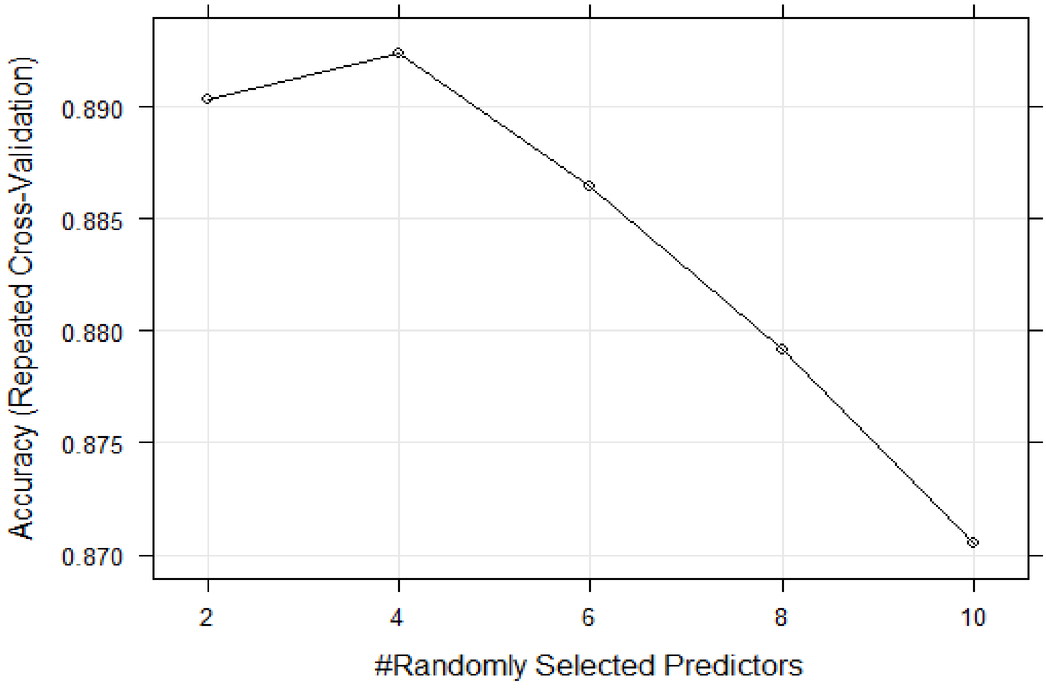
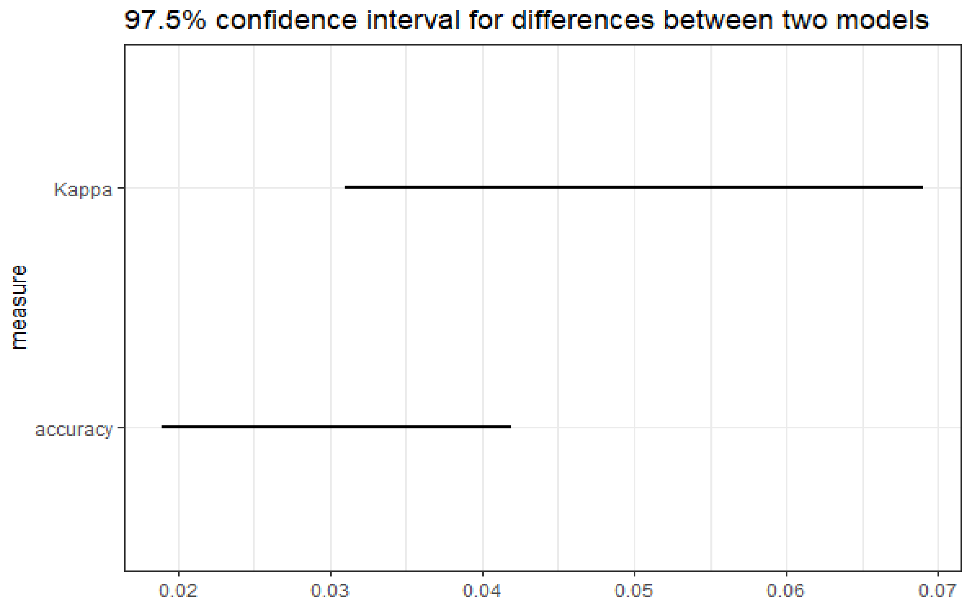


Figure 6: 97.5% confidence interval for accuracy and Kappa


97.5% confidence interval for differences between two models
measure
Kappa
accuracy