Predicting Going Concern of Companies Using Text Mining and Data Mining Approaches

Hamid Abbaskhani, Asgar Pakmaram, Nader Rezaei*
Department of Accounting, Bonab Branch, Islamic Azad University, Bonab, Iran

Jamal Bahri Sales
Department of Accounting, Urmia Branch, Islamic Azad University, Urmia, Iran

Abstract

The linguistic features of the information provided by the business unit can facilitate the achievement of the objectives of conveying economic facts. Thus, in recent years, such features have always been taken into account in accounting and behavioral finance studies. Therefore, the purpose of this study is to determine the ability to predict the going concern of companies using structured and unstructured data, as well as any changes that occur in it because of adding unstructured variables to purely data mining models. In addition, if the results are different, are the difference significant or non-significant? The study period was from 2012 to 2021 and the study sample included 540 company years listed on Tehran Stock Exchange. The tone of the auditor's report was measured using the Mayew et al. (2015) and the Visvanathan (2021) models. The MAXQDA 20 text analysis software and the Loughran and MacDonald (2015) dictionary were also used to process the data. Data analysis and hypothesis testing were done using the logit regression model and the Vuong test. The results of the test of the first hypothesis indicate that the text-based method model has a higher coefficient of determination than the data-based method model, and the test of the second hypothesis shows that there is a significant difference in the exponential explanatory power of the data-based method model and the data-based method model in companies.

Keywords: Auditor Report, Sentiment Analysis, Going Concern, Text Mining

1. Introduction

Predicting bankruptcies has been an ongoing research topic in the accounting and financial fields since the late 1960s. Many researchers developed a more robust bankruptcy-forecasting model for the accuracy of classification. While early studies adopted statistical techniques such as multiple resolution analysis (Altman, 1968) and logit analysis (Hamer, 1983; Ohlson, 1980), later studies adopted artificial intelligence approaches such as artificial neural networks (Lashno & Spector, 1996; Odom & Sharda, 1990), decision trees (Shaw & Gentry, 1990), and support vector machines (Shin et al., 2005) as alternative methods for business prediction problems.

The bankruptcy of companies is related to the financial situation of the company and the external economic situation. Despite ongoing research on the construction of bankruptcy prediction models in terms of modeling techniques, such as statistical methods and artificial intelligence techniques, studies on the use of qualitative information for a bankruptcy prediction model have not yet been conducted. Although the use of financial ratios to model bankruptcy prediction is insufficient, research on the development of bankruptcy prediction models mainly uses superior financial ratios as input variables.

Bankruptcy forecasting models based only on financial ratios have several limitations. Accounting information, such as financial ratios, is based on historical data and is usually determined one year before bankruptcy. The bankruptcy prediction model based on financial ratios is considered a static model (Altman et al., 2010). There is a time interval between the endpoint of the financial statements and the credit rating point. In addition, fiscal ratios do not take into account environmental factors such as external economic situations. The use of financial ratios alone may not be sufficient to construct a bankruptcy-forecasting model, as they do not reflect the latest information and essentially reflect the company's past internal accounting information. To complement the accounting information, qualitative data needs to be added to the standard bankruptcy-forecasting model.

Some past studies have attempted to use non-financial information other than internal accounting information, such as the type of business, firm age, and the number of employees (Altman et al., 2010; Grunert et al, 2005; Pervan & Kuvek, 2013), but these efforts continue to merely reflect the non-financial internal information of a company due to a lack of technology to obtain and analyze qualitative information produced from an external source. Today, vast amounts of data, including news, blogs, and social networking services, are available on the Web. With the increasing volume of

unstructured textual data, big data analysis techniques, particularly text mining, have received considerable attention both in academia and in industry. However, research on the impact of qualitative information on the forecasting model is still in its infancy and is limited to specific applications such as stock forecasting. Therefore, big data analysis techniques, such as text mining, need to be used for various business forecasting issues, including credit risk assessment.

Bankrupt companies in different stages of bankruptcy have poor financial performance (Campbell et al., 2008). Therefore, the importance of bankruptcy and identifying important and effective factors for it is clear and obvious; And despite the fact that so far many researches have examined the quantitative dimensions of financial reports and information to predict the bankruptcy of companies, this research tried to investigate the effect of the qualitative dimensions of financial reports in predicting bankruptcy. The meaning of tone in financial reports is its positive or negative degree (Mirali et al., 2018). ~~In other words, researchers first quantify the tone by using special criteria and tools, then evaluate the extent and manner of its influence on the desired factors(Rahnamay Roodposhti et al., 2012).~~ Therefore, by using certain methods and tools of text mining, the tone of the text is converted into a quantitative state, and then with the selected model, its impact on the desired factors is evaluated (Siano & Wysocki, 2021).

2. Theoretical Foundations and Research Background

2.1. Theoretical

~~Bankruptcy is an unpleasant event that, on the one hand, companies exposed to bankruptcy face a severe decrease in market value, and on the other hand, managers and beneficiaries are severely affected by the negative effects of bankruptcy (Aref Manesh & Bazrafshan, 2015). The reasons for bankruptcy can be classified into two categories, internal and external. Internal reasons are often caused by wrong decisions by the business unit, and external reasons include the characteristics of the economic system, competition, business fluctuations, etc (Khajavi & Amiri, 2012).~~ Nowadays, country is faced with increasing downward pressure on its economy, along with an expanding business risk on listed companies. Listed companies, as the solid foundation of the national economy, once they face a financial crisis, will experience hazards from multiple perspectives. Therefore, the construction of an effective financial crisis early warning model can help beneficiaries predict risks (Zhang et al., 2022). beneficiaries often look for ways to predict corporate bankruptcy. Therefore, the need for information, especially

qualitative information, along with the quantitative information published by the company, has received the attention of beneficiaries more than in the past. Writings in financial reports can focus on persuasiveness. One of the important methods of persuasion is reflecting and repeating certain words of information in the text, and this method emphasizes the tone of information disclosure (Henry, 2008).

Ideas and thoughts are reflected in the tone of the messages in annual reports (Huang et al., 2013; Yekini et al., 2016). ~~In qualitative texts, the tone of the text refers to the use of positive terms in contrast to negative terms (Mohseni & Rahnamay roodposhti, 2020).~~ In qualitative texts, positive words are used against negative words to evaluate the tone of the text (Kou, J., 2022). The pessimistic tone of financial statements will cause investors to respond negatively (Feldman et al., 2010; Loughran & McDonald, 2011). Previous research has linked the tone of financial reporting to the company's economic performance and business risk.

Loughran and McDonald (2011) found that words that have a negative tone are more effective and reliable than positive words. This view is in line with the results of Law and Mills' (2015) psychological research because humans tend to process more negative information than positive information. Another study found that a pessimistic tone influences readers' decisions in a statistically meaningful way (Garcia et al., 2013).

To measure the tone of writing in managers' reports, researchers used a variety of methods. There are two common approaches to content analysis: the first is based on counting the frequency of specific words (dictionary), and the other is a machine learning classification algorithm method based on assigning an experimental data set to specific categories using a manual coding mechanism (Kashanipoor et al., 2020). In financial research, the methodology based on counting the frequency of specific words is more common and assigns words to different classifications based on predefined rules (Loughran & McDonald, 2011).

In this research, a method based on counting the frequency of specific words was used. There is no general consensus in the research literature on text tone word lists, but two lists of words provide the most appropriate classification for use in text analysis (Davis et al., 2015). The first list includes the Loughran and McDonald (2011) dictionary, which is specifically designed to analyze the text of financial and accounting reports, while the second list, the Mohammad & Turney (2013) dictionary, contains the general word list (Mohammad et al., 2019).

This study broadly refers to the information literature of the Annual Reports (Brown & Taker, 2011; Cole & Jones, 2005; Feldman et al., 2010), the bankruptcy prediction literature (Altman, 1968; Beaver et al., 2005; Ohlson, 1980; Shumway, 2001; Zmijewski, 1984), and the literature that studies the auditor's boundaries for its going concern (Carson et al. 2012). This study also contributes to the growing literature on the importance of qualitative disclosure using automated language techniques (Tetlock, 2007; Tetlock et al., 2008; Li, 2010) and in particular fills the gap identified by Li (2011) that linguistic analysis may be useful for predicting bankruptcy (Mayew et al., 2015).

Two important research issues raise researchers' interest in giving investors early warning signals through auditor disclosure. Firstly, does the tone of the auditor's report and the business unit's going concern disclosures help predict whether a business will continue to operate? Secondly, to what extent are the outcomes of the first question different from the purely structured data? Discussing the advantages of extending the textual disclosure of financial statements can be aided by the answers to these questions.

2.2. Research Background

Mayew et al. (2015) examined the role of textual disclosure in a firm's financial statements to predict a firm's ability to continue as a going concern. Using a sample of 262 firms that filed for bankruptcy over the period 1995-2011 and a matched set of control firms, they find that both the management's opinion and the textual features of management discussion and analysis disclosure together provide significant explanatory power in predicting whether a firm will cease to exist as a going concern. In addition, the ability to predict MD&A disclosure is incremental to financial ratios, market-based variables, and the auditor's opinion. The most important finding of the study is that information in MD&A disclosure is more useful in predicting bankruptcy three years before it occurs. This indicates that MD&A disclosures are more timely than financial ratios, making them a leading indicator of going concern problems.

According to Jo and Shin (2016), qualitative information should be added to the conventional bankruptcy prediction model to complement accounting information. This study proposes a bankruptcy prediction model for small and medium-sized Korean construction companies using both quantitative data such as financial ratios and qualitative data obtained from economic news articles. The performance of the proposed method depends on how well the types of information are converted from qualitative to quantitative information suitable for incorporating into the bankruptcy prediction model. In addition, big data analysis techniques, especially text

mining, have been used to process qualitative information. The proposed method involves analyzing keyword-based sentiment analysis using a domain-specific sentiment lexicon to extract sentiment from economic news articles. Experimental results showed that combining qualitative data based on extensive data analysis in the traditional model of bankruptcy forecasting based on accounting information effectively increases forecasting performance. The experimental results showed that incorporating qualitative information based on big data analytics into the traditional bankruptcy prediction model based on accounting information is effective for enhancing predictive performance. The sentiment variable extracted from economic news articles impacted corporate bankruptcy. In particular, a negative sentiment variable improved the accuracy of predicting corporate bankruptcy because the corporate bankruptcy of construction companies is sensitive to poor economic conditions.

Lopatta et al. (2017) examine whether the language used in 10-K filings reflects a firm's risk of bankruptcy. They use propensity score matching to find healthy matches. Based on a logit model of failing and vital firms, their findings indicate that firms at risk of bankruptcy use significantly more negative words in their 10-K filings than comparable vital companies. They confirm the findings of previous accounting and finance research with their investigation. Beyond the reported financials, 10-K filings contain valuable information. Additionally, they show that 10-Ks filed in the year of a firm's collapse contain a higher proportion of litigious words than healthy businesses. This indicates that the management of failing firms is already dealing with legal issues when reporting financials before bankruptcy. Their results suggest that analysts ought to include the presentation of financials in their assessment of bankruptcy risk as it contains explanatory and predictive power beyond the financial ratios.

Dey et al., (2017) report that due to the vast amount of textual information generated across various sources on the web, they have begun to combine relevant structured and unstructured data to improve predictions. This study provides a generic deep-learning framework for predictive analysis using structured and unstructured data. They also offer a case study to validate the performance and application of the proposed framework in which LSTM is used to predict the movement direction of structured data utilizing events extracted from news articles. Experimental results show that the proposed model outperforms the existing baseline.

Li and Wang (2018) conducted a study in which they compared statistical and machine learning (ML) methods for predicting bankruptcy using Chinese listed companies. They began by selecting

the most appropriate indicators using statistical methods. Different indicators may have different characteristics, and not all indicators can be analyzed. The indicators will be more convincing after filtering the data. Unlike previous research methods, researchers used the same sample set to conduct their experiments. The result proves the effectiveness of the machine learning method. Furthermore, with 95.9 percent accuracy, the test outperforms previous studies.

By creating a comprehensive corporate failure-related lexicon, Elsayed et al. (2020) explored the incremental explanatory power of narrative-related disclosures in predicting corporate failure. They found that corporate failure-related narrative disclosures significantly predict firms' failure up to two years ahead of actual failure. Additionally, they found that a financially distressed firm would become more vulnerable when financial constraints befall, which in turn would precipitate corporate failure. Various robustness tests assured the credibility of the explanatory ability of corporate failure-related narrative disclosures to predict corporate failure. Collectively, their results showed the feasibility of these narrative-related disclosures in improving the explanatory power of models that predict corporate failure.

According to Gutierrez et al. (2020), investors, regulators, and academics question the usefulness of going concern opinions (GCOs). They assessed whether GCOs provide incremental information, relative to other predictors of corporate default. Their measure of incremental information was the additional predictive power that GCOs give to a default model. Utilizing data from 1996 to 2015, initially, they found no difference in predictive power between GCOs alone and a default model that includes financial ratios. However, there was an imperfect overlap between GCOs and other predictors. They showed that GCOs increase the predictive power of several models that include ratios, market variables, probability of default estimates, and credit ratings. Using a model that includes ratios and market variables, GCOs increased the number of predicted defaults by 4.4%, without increasing Type II errors. Their findings suggested that GCOs summarize a complex set of conditions not captured by other predictors of default.

Lohmann and Ohliger (2020) say the structural and linguistic characteristics of companies' annual reports (e.g., their length, complexity, and linguistic tone) and the qualitative information they contain (e.g., on the risks a company potentially faces) provide useful insights that can help increase the accuracy of predicting bankruptcy. They use a sample of German companies that they compiled through propensity score matching to examine what type of textual information allows them to discriminate accurately between

companies that are likely to go bankrupt and companies that, although financially distressed, are likely to remain solvent. Their findings provide empirical evidence that both the structural and linguistic characteristics of annual reports and the qualitative information they contain help discriminate between effectively bankrupt companies and companies that are solvent but financially distressed. Furthermore, the study provides empirical evidence that the "management obfuscation hypothesis" is valid because the tone of annual reports produced by bankrupt companies is quantifiably less negative than that of reports produced by companies that, although financially distressed, are likely to remain solvent.

~~Toorchi and lari dashtebayaz (2021) conducted a study by considering qualitative criteria along with quantitative criteria to predict bankruptcy. With the data of 476 company years, they concluded that the tone of the report of the board of directors can predict the bankruptcy of companies.~~

~~Bozorg Asl et al. (2021) studied the effect of the tone of financial reporting on audit fees. The results of the hypothesis test indicate that the tone of financial reporting has a negative and significant effect on audit fees. The obtained results show that the tone of financial reporting reflects the criteria that auditors consider in assessing audit risk.~~

Visvanathan (2021) in his study explores the role of deferred tax valuation allowances, management's discussion of the ability to continue as a going concern, and auditor going concern opinions in predicting the financial distress of a firm. His study is in line with the development of Mayew et al.'s (2015) analysis by including deferred tax valuation allowances in their framework. To the extent valuation allowances incorporate managers' private information about future profitability, valuation allowances are useful in identifying the transitory nature of losses and thus the going concern status of the firm. Using a sample of firms that filed for bankruptcy over the period 2002–2018, the study shows that increases to valuation allowances are incrementally informative in predicting a firm's ability to continue as a going concern after considering management's textual disclosures, linguistic tone of the MD&A, auditor's going concern opinions, financial statement ratios, and market-based variables.

Nießner et al. (2022) conducted a study by considering qualitative criteria along with quantitative criteria to predict bankruptcy. they concluded that qualitative information of companies' financial statements provides useful information that can increase the accuracy of bankruptcy prediction models.

Zhao et al. (2022) conducted a study in which, in addition to financial features, they proposed a novel framework that combines sentiment tone

features extracted from management discussion and analysis, and financial statement notes to predict financial distress. They found that financially distressed companies were more likely to have weak sentiment. They recommend incorporating sentiment tone features with financial features, as they contribute to predictive performance improvements of all models using only financial features. Economic benefits analysis shows that the proposed framework can correctly identify financially distressed companies.

In the domestic and foreign backgrounds of the research, the bankruptcy and the going concern, and the tone of the auditors' report have been investigated, but those investigations have not been carried out simultaneously or in the economic environment of Iran with internal data. In this research, the researchers have conducted research by considering the cases of the going concern, the tone of the auditors' report, and the positive and negative words of the reports for the domestic companies. As a result of this research, the impact of qualitative and textual data along with the quantitative data of the company is determined for the interested parties.

The following summarizes the contribution of this paper to the development of the research literature.

1- The use of qualitative data in addition to quantitative data improves estimates and forecasts about the company.

2- It fosters the attitude that qualitative data can be used to predict a company's going concern, and it draws more attention to qualitative data in the field of going concern predictions.

3- The present research arouses the interest of researchers to study more in this field and use other qualitative data.

3. Research Hypotheses

The two hypotheses of the present study are:

3.1. The ability to predict the going concern of companies with/without growth opportunities using the text mining approach is greater than the data mining approach.

3.2. The ability to predict companies with/without growth opportunities using the text mining approach is significantly different from expecting the going concern using the data mining approach.

4. Research Method

The research is applied in terms of purpose because its results can be used by potential and actual investors, as well as other groups, and it is correlational in nature because it examines the relationships

between variables using regression analysis. The necessary information on the research literature and theoretical foundations was obtained from library sources, scientific databases, and national and foreign articles. Tehran Stock Exchange Organization and Rahavard Novin software database were utilized to collect research data, reports, and announcements published in the codal network. The MAXQDA software version (2020) and the Loughran and MacDonald (2015) dictionary were then used to perform the text analysis section processing. EViews software version 10 was used to test the hypotheses after extracting the research's numerical and textual data.

4.1. Statistical Population and Statistical Sample

The research statistical population consists of companies listed on the Tehran Stock Exchange and the study period is from 2012 to 2021. In this study, a statistical sample was performed by the systematic elimination method using Article 141 of the Commercial Code to select 27 bankrupt companies and the Q-Tobin ratio to select 27 successful going concern companies. The number of samples used in this research is 540 company years.

4.2. Text Mining Process

The text mining process involves steps according to Figure 1 to extract data from the document (Kumar and Bhatia, 2013).

Figure 1. Text Mining Process

The present study uses the latest updated version of the Loughran and MacDonald dictionary (2015), which is available through the relevant site and contains 354 positive words and 2355 negative words. The translation of this dictionary was used to analyse the contents of the auditor's annual report on the activities and general status of the company. For example, with the assistance of content analysis software, the number of positive words (such as desirable, excellent, and profit) and negative words (unfavorable, weakness, and loss) can be counted in accounting narrations.

The frequency of positive and negative words is reflective of the tone of the language. We measure the auditor's going concern statement using an index variable (GC_AUD) and if the auditor is

unsure of the company's going concern, its value becomes zero; otherwise, it becomes one (Mayew et al., 2015).

The audit reports of the sample companies were extracted from the codal site by the researchers for this section. They then entered the Maxqda software to determine the word count within each report. Afterward, they counted the number of positive and negative words in auditor reports using the Loughran and McDonald dictionaries. For the index variable (GC_AUD), Standard No. 570 of Auditing Standards entitled going concern has been used. A company that is experiencing going concern difficulties has two instances (although only one is stated in the standard) of signs referred to in this standard that suggest a serious doubt as to whether going concern exists.

4.3. Logit Model

The dependent variable in this model is a two-state variable equal to the logarithm of the probability that a particular event (bankruptcy) will happen. The linear probability model as equation 1 can be written in the form of a logistic regression function as equation 2.

Equation 1) $$Y = b_1 + b_2 X_i$$

Equation 2) $$Ln\left(\frac{p}{1-p}\right) = b_1 + b_2 X_i + m$$

Therefore, the probability of an event occurring is described in equation 3.

Equation 3) $$p = \frac{1}{1 + e^{-(b_1 + b_2 X_i)}}$$

The maximum probability method is used to estimate Equation 5. We take zero to represent bankruptcy. If the result is greater than 0.5 decimal places (which is used for the company's equal index of bankruptcy or non-bankruptcy), the company has a lower chance of continuing as a going concern. Researchers who have used this method include Mayew et al. (2015) and Lee and Wang (2018).

4.4. Vuong Z Test

For comparison of the power of two models in a common statistical sample. It is necessary to take into account the coefficient of determination obtained from the estimation of the two models. Because the power of the model is determined by the amount of the coefficient of determination. A model with a higher determination coefficient is more power in explaining and forecasting the

dependent variable. However, whether the difference in the coefficients of determination of the two models is statistically significant or not, a test must be carried out. The desired test for comparing the difference between the coefficients of determination of both models was introduced by Vuong (1989), known as the Vuong Z test (Banimahd et al., 2016: 199).

5. Research Variables and Models

5.1. Dependent Variable

Bankruptcy: In Iran, Article 141 of a 1968-approved amending bill to a section of the Commercial Law serves as the foundation for bankruptcy. According to this article, the board of directors is required to summon an extraordinary general meeting of shareholders as soon as at least half of the company's capital is lost as a result of losses so that the issue of the company's survival or liquidation can be discussed and voted on.

5.2. Independent Variables

Research variables are categorized into both quantitative and qualitative. The quantitative variables are retrieved from the financial statements. The qualitative variables were collected by counting the positive and negative words and dubious phrases from the going concern in the auditor's report.

In this study, 11 independent variables were used, which are presented in Table (1) and have been used in past studies by national and international researchers.

Table 1. Research Variables

independent variables	***researchers***	***symbols***
retained earnings to total assets ratio	*Mayow et al. (2015), ~~Hajiha and babaei (2016),~~ Li and Wang (2018), ~~Bayat et al. (2018),~~ Rowland et al. (2021)*	*Reta*
net profit to total assets ratio	*~~Hajiha and babaei (2016)~~, Elsayed & Elshandidy (2020), ~~Saroei et al. (2020)~~ Rowland et al. (2021)*	*Neta*
operating profit to total assets ratio	*Mayow et al. (2015), Li and Wang (2018), ~~Saroei et al. (2020)~~, Rowland et al. (2021)*	*Ebitta*
current assets to current liabilities ratio	*~~Hajiha and babaei (2016),~~ Li and Wang (2018),~~ Bayat et al. (2018),~~ Elsayed & Elshandidy (2020), Rowland et al. (2021)*	*Cacl*
working capital to total assets ratio	*Mayow et al. (2015), ~~Saroei et al. (2020)~~ Rowland et al. (2021)*	*Wcta*

Total liabilities to total assets ratio	*Li and Wang (2018), ~~Bayat et al. (2018),~~ Rowland et al. (2021)*	*Tlta*
sale revenue to total assets ratio	*Mayow et al. (2015), ~~Bayat et al. (2018),~~ Saroei et al. (2020), Rowland et al. (2021)*	*Saleta*
growth opportunity	*Namazi et al. (2018)*	*growth*
Positive words	*Wang et al. (2013), Mayow et al. (2015), Jo and Shin (2016), Visvanathan (2021)*	*Posmda*
Negative words	*Wang et al. (2013), Mayow et al. (2015), Jo and Shin (2016), Visvanathan (2021)*	*Negmda*
Expressing the substantial doubt of the auditor	*Mayow et al. (2015), Visvanathan (2021)*	*Gc_aud*

5.3. Research model

According to the studies of Mayow et al. (2015) and Visvanathan (2021), the research model is based on research hypotheses and a data mining approach as equation 4.

Equation 4)
$$\Pr(BRUPT_{t+1}) = \beta_0 + \beta_1 RETA_t + \beta_2 NETA_t + \beta_3 EBITTA_t + \beta_4 CACL_t + \beta_5 WCTA_t + \beta_6 TLTA_t + \beta_7 SALETA_t + \beta_8 Growth_t + \vartheta_t$$

And the research model is based on research hypotheses and text analysis approach as equation 5.

Equation 5)
$$Pr(BRUPT_{t+1}) = \beta_0 + \beta_1 RETA_t + \beta_2 NETA_t + \beta_3 EBITTA_t + \beta_4 CACL_t + \beta_5 WCTA_t + \beta_6 TLTA_t + \beta_7 SALETA_t + \beta_8 Growth_t + \beta_9 POSMDA_t + \beta_{10} NEGMDA_t + \beta_{11} GC\_AUD_t + \vartheta_t$$

6. Analysis of Research Data and Findings

5.4. Unit Root Test

Dummy regression occurs when nonstationary variables are present in the model. The test presented by Levin et al. (2002: 5) was used to evaluate the significance of the variables. When the time dimensions are large enough, this test is more efficient and has more power than other static tests (Najafzadeh et al., 2022). Table (2) shows the results of a test of the reliability of independent research variables.

Table 2. Test of Reliability of Independent Research Variables

variables	***Levin, Lin & Chu test statistics***	***Significance level(prob.)***

Reta	*-4.34626*	*0.0000*
Neta	*-7.21567*	*0.0000*
Ebitta	*-6.81778*	*0.0000*
Cacl	*-8.16295*	*0.0000*
Wcta	*-7.80360*	*0.0000*
Tlta	*-6.89345*	*0.0000*
Saleta	*-10.8886*	*0.0000*
growth	*-7.31375*	*0.0000*
Posmda	*-7.84598*	*0.0000*
Negmda	*-9.04359*	*0.0000*
Gc_aud	*-7.44742*	*0.0000*

If the variables are nonstationary, the co-integration method is used to allow the original values of the variables to be used while ensuring that the regression results are not a dummy. If one of their linear combinations is stationary, a set of values is said to be co-integrated. Therefore, if the explanatory and dependent variable processes co-integrate in a regression model, the possibility of dummy regression is eliminated (Banimahd et al., 2016: 184). The unit root test results in Table (3) show that the distribution of error values in both models in Table (4) is significant. As a result, the explanatory and dependent variables' linear relationships are co-integrated.

Table 3. Unit Root Test Results of Error Values of Regression Models

Variables	***Levin, lin & chu test statistics***	***Significance level(prob.)***
Error-values of data mining regression model	*-21.7842*	*0.0000*
Error-values of text mining regression model	*-24.2126*	*0.0000*

5.5. Results of Fitting the Regression Models of the Research

Table (4) compares the results of Logit regression equation estimation to make it easier to compare the explanatory ability of data mining and text mining models for bankruptcy prediction. In the data mining technique, the logit regression model has a coefficient of determination of 62 percent, while in the text mining approach, it has a coefficient of determination of 64 percent.

Table 4. Results of Estimating Dependent Variable Hypotheses

Variables	***Data mining model***	***Text mining model***

C	*-0.26*	*-0.99*
Reta	*5.31***	*5.27****
Neta	*-4.68*	*-6.38*
Ebitta	*0.66*	*2.24*
Cacl	*-3.16****	*-3.29****
Wcta	*2.34****	*2.48****
Tlta	*-0.76***	*-0.74***
Saleta	*-0.49****	*-0.53****
Growth	*0.63**	*0.58**
Posmda		*-0.25*
Negmda		*0.45**
Gc_aud		*1.13****
R-squared	*0.62*	*0.64*

Symbols ***, ** and * indicate significance levels of 99%, 95% and 90%, respectively.

These coefficients indicate that the presence of 3 qualitative variables, positive words, negative words, and the auditor's expression of doubt, along with quantitative model variables, it improves the explanatory power of the model. Although the variable of positive words is not significant, but the variables of negative words and the auditor's expression of doubt with a significance level of 0.1 and 0.01 play an essential role in predicting bankruptcy in the text mining model.

5.6. Results of Vuong Test of Research Models

The Vuong Z statistic was used to ensure that the incremental explanatory power of the text-mining model compared to the data-mining model in companies with/without growth opportunities was different in Table (5). Consequently, the incremental explanatory power of a model with a bigger R^2 is greater. The text mining model has more explaining power than the data mining model in enterprises with and without growth opportunities. Overall, these results show that the increasing explanatory power of the text mining model in companies with and without growth opportunities differs significantly from the data mining model.

Table 5. Vuong Test Results of Models

Vuong statistic value	***z statistics***	***Significance level(prob.)***	***Test result***
-2.5476	*0.01*	*0.05*	*The hypothesis is confirmed*

7. Conclusions and Suggestions

Text mining is the process of obtaining high-quality information from unstructured or semi-structured texts or data (Marti, 2003). Accordingly, in the areas of modern accounting and behavioral finance, particular attention has been paid to the relationship between the linguistic characteristics of enterprises' annual reports and their behavior and economic results (Davis et al., 2012; Hong et al., 2014). In recent years, the study of linguistic features of financial reporting in experimental accounting research has been prompted by the variety of disclosable issues, the diversity of different industries of international companies, and the existence of various institutions that formulate accounting standards at the global level. Although numerous studies have been conducted on bankruptcy and the influence of different factors in its determination, the influence of the tone of the financial reports as a linguistic characteristic of the company's financial reports was not considered. It was found that adding textual variables to data mining models with the presence of company size improves the coefficient of determination of logit regression models, according to the results obtained from testing the first hypothesis of the research in Table (4). Moreover, the predictive power of text mining is greater, as shown in Table (5), and the difference is significant.

Finally, the search hypotheses are confirmed basis on the results obtained and the corresponding coefficients. Furthermore, the findings of this study are consistent with those of Mayow et al. (2015), Al-Sayed et al. (2020), Lohmann & Ohliger (2020), Viswanathan (2021), and ~~Hosseini and Jamalianpour (2022)~~ Nießner et al. (2022). It is suggested that the Auditing Organization, Tehran Stock Exchange and Securities Organization design a specific framework that includes compiling explanatory reports with a specific lexicon for both formulating new laws and amending previous cases with the knowledge of how auditors manage perception. In addition, auditing firms are urged to take into account the tone of financial statements in assessing the level of risk to the client company, the planning of operations, and the volume of audit tests, among other factors.

Ultimately, it is suggested that the following be investigated in future research:

1. Comparing the predictive power and going concern of companies using text mining and data mining approaches with qualitative variables of the activity report of the board of directors
2. The effect of financial reporting tone on the comparability of financial statements.

3. Further, variables in the current study were examined over 10 years and in a sample of 54 companies, which suggests a longer period with a much larger sample can be useful.

In the research process, there are a set of conditions and cases that are out of control but can potentially affect the results. it is necessary to examine the results of the research, taking into account the existing limitations. The limitations of this study were as follows:

1. Lack of Persian dictionaries that can be used as a standard tool to measure the tone of writing in the field of financial research. Thus, due to the use of English dictionaries in translation and the linguistic differences, if there were a standard dictionary in Persian, the reliability of research tools would increase.

2. As the word file of the auditor's reports was unavailable, calculating the financial reporting tone index was very challenging.

3. According to the entire descriptions of the text in the report, the results of the research are obtained in the tone inferred. Nonetheless, the tone inferred by the investor from a fraction of the text differs from the tone inferred from the entire text because there is no guarantee of an equal distribution of positive, negative, or neutral words in all paragraphs.

References:

Altman, E.I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *Journal of Finance*, 23, 589-609.

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, 95-127.

~~Aref Manesh, Z., & Bazrafshan, A. (2015). Earnings behavior in bankrupt firms: the role of auditor. *Journal of Asset Management and Financing*, *2*(4), 1-14. (In Persian)~~

Banimahd, B., Arabi, M. & Hasanpor, Sh. (2016). Experimental research and methodology in accounting, Tehran, Termeh Publications.

~~Bozorg Asl, M., marfo, M., & mahannejad, M. (2021). The Effect of Financial Reporting Tone on Audit Fees of Listed Companies in Tehran Stock Exchange. *Empirical Studies in Financial Accounting*, *18*(72), 79-107. (In Persian)~~

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63* (6), 2899-2939.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. Review of Accounting Studies, 20(2), 639–673.

Dey, L., Meisheri, H., & Verma, I. (2017). Predictive Analytics with Structured and Unstructured data - a Deep Learning based Approach. *IEEE. Informatics Bull*., 18, 27-34.

Drass, K. A. (2019). Text Analysis and Text-Analysis Software: A Comparison of Assumptions. *New Technology in Sociology*, 155–162.

Elsayed, M., Elsayed, M., Elshandidy, T., & Elshandidy, T. (2020). Do narrative-related disclosures predict corporate failure? Evidence from UK non-financial publicly quoted firms. *International Review of Financial Analysis*, *71*, 101555.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. Review of Accounting Studies, 15(4), 915–953.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), 1267–1300.

Grunert, J.P., Norden, L., Weber, M., & Weber, M. (2005). The role of non-financial factors in internal credit ratings. *Journal of Banking and Finance*, 29, 509-531.

Gutierrez, E., Krupa, J., Minutti-Meza, M., & Vulcheva, M. (2020). Do going concern opinions provide incremental information to predict corporate defaults?. *Review of Accounting Studies*, *25*(4), 1344-1381.

Hamer, M.M. (1983). Failure prediction: sensitivity of classification accuracy to alternative statistical methods and variable sets. *Journal of Accounting and Public policy*, 2, 289-307.

Henry, E. (2008). Are investors influenced by how earnings press releases are written?. *The Journal of Business Communication (1973)*, *45*(4), 363-407.

Hobbs, J. R., Walker, D. E., & Amsler, R. A. (1982). Natural language access to structured text. *Proceedings of the ninth Conference on Computational Linguistics* -.

Hotho, A., Nürnberger, A., & Paass, G. (2005). A brief survey of text mining. *LDV Forum*, 20, 19-62.

Huang, X., Teoh, S. H., & Zhang, Y. (2013). Tone management. *The Accounting Revie*w, 89(3), 1083–1113.

Jo, N., & Shin, K. (2016). Bankruptcy prediction modeling using qualitative information based on big data analytics.

~~Khajavi, S., & Amiri, F. S. (2012). Recognition of Efficient Factors Affecting in companies' bankruptcy using TOPSIS_AHP. *Empirical Studies in Financial Accounting*, *10*(38), 69-90. (In Persian)~~

Kashanipoor, M., Aghaee, M.A., & Mohseni Namaghi, D. (2020). Information Disclosure Tone and Future Performance. *Accounting and Auditing Review*, 26(4), 570-594. (In Persian)

Kou, J. (2022). *Analysing Housing Price in Australia with Data Science Methods* (Doctoral dissertation, Victoria University).

Kumar, L., & Bhatia, P. (2013). Text mining: concepts, process, and applications. *Journal of Global Research in Computer Sciences*, 4, 36-39.

Law, K. K., & Mills, L. F. (2015). Taxes and financial constraints: Evidence from linguistic cues. *Journal of Accounting Research*, 53(4), 777–819.

Li, Y., & Wang, Y. (2018). Machine learning methods of bankruptcy prediction using accounting ratios. *Open Journal of Business and Management*, 06, 1-20.

Leshno, M., & Spector, Y. (1996). Neural network prediction analysis: the bankruptcy case. *Neurocomputing*, 10, 125-147.

Levin, A., Lin, C.F., & Chu, C-S.J. (2002). Unit root test in panel data. *Journal of Econometrics*, 108, 1-22.

Lohmann, C., & Ohliger, T. (2020). Bankruptcy prediction and the discriminatory power of annual reports: empirical evidence from financially distressed German companies. *Journal of Business Economics*, *90*(1), 137-172.

Lopatta, K., Gloger, M. A., & Jaeschke, R. (2017). Can language predict bankruptcy? The explanatory power of tone in 10-K filings. *Accounting Perspectives*, *16*(4), 315-343.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65.

Marti Hearst: What is Text Mining? (https://people.ischool.berkeley.edu/~hearst/text-mining.html)

Mayew, W.J., Sethuraman, M., & Venkatachalam, M. (2015). MD&A Disclosure and the Firm's Ability to Continue as a Going Concern. *The Accounting Review*, 90(4), 1621-1651.

Merkl-Davies, D. M., Brennan, N. M. (2007). Discretionary Disclosure Strategies in Corporate Narratives: Incremental Information or Impressions Management? *Journal of Accounting Literature*: 27, 116- 196.

Mirali, M., Gholami moghaddam, F., Hesarzadeh, R. (2018). Investigation of the Relationship between Financial Reporting Tone with Future Corporate Performance and Market Return. *Financial Accounting Knowledge*, 5(3), 81-98. (In Persian)

Najafzadeh, A., Farzinvash, A., Yosefi Sheikhrobat, M. & Naser, A. (2021). Asymmetric behavior of the effectiveness of fiscal policies in the smooth transition process, Tehran, Mofid University.

Nießner, T., Gross, D. H., & Schumann, M. (2022). Evidential Strategies in Financial Statement Analysis: A Corpus Linguistic

Text Mining Approach to Bankruptcy Prediction. *Journal of Risk and Financial Management*, *15*(10), 459.

Odom, M.D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *Joint international conference on neural networks*, 2,163-168.

Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 18, 109-131.

Pervan, I., & Kuvek, T. (2013). The relative importance of financial ratios and nonfinancial variables in predicting insolvency. *Croatian Operational Research Review*, 4, 187-197.

~~Rahnamay Roodposhti, F., Nikoomaram, H., & Nonahal Nahr, A. A. (2012). Evaluating the Effects of Language Judgmental and Cognitive Approaches in Accounting Narratives. *Accounting and Auditing Review*, *19*(2), 47-72. (In Persian)~~

Rowland, Z., Kasych, A., & Suler, P. (2021). Prediction of financial distress: case of mining enterprises in Czech Republic. *Ekonomicko-manazerske spektrum*, *15*(1), 1-14.

Shaw, M.J., & Gentry, J.A. (1990). Inductive learning for risk classification. IEEE expert, 5, 47-53.

Shin, K., Lee, T.S., & Kim, H. (2005). An application of support vector machines in the bankruptcy prediction model. *Expert syst. Appl*., 28, 127-135.

Siano, F., & Wysocki, P. (2021). Transfer learning and textual analysis of accounting disclosures: Applying big data methods to small (er) datasets. *Accounting Horizons*, *35*(3), 217-244.

~~Toorchi, M., & lari dashtebayaz, M. (2021). Tone of Board Activity Report and Bankruptcy Prediction. *Empirical Research in Accounting*, 11(2), 137-158. (In Persian)~~

Visvanathan, G. (2021). Is information in deferred tax valuation allowance useful in predicting the firm's ability to continue as a going concern incremental to MD&A disclosures and auditor's going concern opinions?. *International Journal of Disclosure and Governance*, *18*(3), 223-239.

Yekini, L. S., Wisniewski, T. P., & Millo, Y. (2016). Market reaction to the positiveness of annual report narratives. *The British Accounting Review*, 48(4), 415–430.

Zhang, Z., Luo, M., Hu, Z., & Niu, H. (2022). Textual Emotional Tone and Financial Crisis Identification in Chinese Companies: A Multi-Source Data Analysis Based on Machine Learning. *Applied Sciences*, *12*(13), 6662.

Zhao, Y., Wei, S., Guo, Y., Yang, Q., & Kou, G. (2022). FisrEbp: Enterprise Bankruptcy Prediction via Fusing its Intra-risk and Spillover-Risk. *arXiv preprint arXiv:2202.03874*.