IJAAF; Vol. 8, No. 3, Summer 2024, Serial 28

Ferdowsi University of Mashhad

Iranian Journal of
Accounting, Auditing & Finance

Quarterly

RESEARCH ARTICLE

Enhancing Going Concern Prediction Models: Integrating Text Mining with 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

How to cite this article:
Abbaskhani, H., Pakmaram, A., Rezaei, N., & Bahri Sales, J. (2024). Enhancing Going Concern Prediction Models: Integrating Text Mining with Data Mining Approaches. Iranian Journal of Accounting, Auditing and Finance, 8(3), 27-42. doi: 10.22067/ijaaf.2024.43123.1217
https://ijaaf.um.ac.ir/article_43123.html

ARTICLE INFO

Article History
Received: 2023-09-25
Accepted: 2023-12-10
Published online: 2024-07-06

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

Abstract

The linguistic features embedded within business unit information play a crucial role in effectively conveying economic realities, a consideration increasingly recognized in accounting and behavioral finance research. This study endeavors to assess the predictive capacity of companies' going concern status by integrating structured and unstructured data, while also evaluating the impact of incorporating unstructured variables into traditional data mining models. Spanning from 2012 to 2021, the study encompasses a sample of 540 company years listed on the Tehran Stock Exchange. Tone analysis of auditor reports was conducted using models by Mayew et al. (2015) and Visvanathan (2021), while MAXQDA 20 text analysis software and the Loughran and McDonald (2015) dictionary facilitated data processing. Data analysis and hypothesis testing were performed using the logit regression model and the Vuong test. The findings support the first hypothesis, indicating that the text-based model yields a higher coefficient of determination compared to the data-based approach. Moreover, the second hypothesis reveals a significant discrepancy in the explanatory power between the data-based and integrated text-based models within companies.

https://doi.org/10.22067/ijaaf.2024.43123.1217

NUMBER OF REFERENCES
56

NUMBER OF FIGURES
1

NUMBER OF TABLES
5

Homepage: https://ijaaf.um.ac.ir
E-Issn: 2717-4131
P-Issn: 2588-6142

*Corresponding Author: Nader Rezaei
Email: nader.rezaei@bonabiau.ac.ir
Tel: 09144815093
ORCID: 0000-0001-9907-4648

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 classification accuracy. 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 (Leshno and Spector, 1996; Odom and Sharda, 1990), decision trees (Shaw and Gentry, 1990), and support vector machines (Shin et al., 2005) as alternative methods for business prediction problems.

A company's bankruptcy 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 using 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 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. Using 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. Qualitative data needs to be added to the standard bankruptcy forecasting model to complement the accounting information.

Some past studies have attempted to use nonfinancial 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 and Kuvek, 2013), but these efforts continue to merely reflect the nonfinancial 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 online. With the increasing volume of unstructured textual data, big data analysis techniques, particularly text mining, have received considerable attention in academia and 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 obvious; even though so far, many researchers 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). Therefore, using certain methods and tools of text mining, the text's tone is converted into a quantitative state, and then with the selected model, its impact on the desired factors is evaluated (Siano and Wysocki, 2021).

2. Theoretical foundations and research background

2.1 Theoretical

The country faces increasing downward pressure on its economy and an expanding business risk on listed companies. Listed companies, as the solid foundation of the national economy, will experience hazards from multiple perspectives once they face a financial crisis. Therefore, constructing 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, which 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., 2014; Yekini et al., 2016). In qualitative texts, positive words are used against negative words to evaluate the text's tone (Kou, 2022). The pessimistic tone of financial statements will cause investors to respond negatively (Feldman et al., 2010; Loughran and 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 aligns with Law and Mills' (2015) psychological research results because humans tend to process more negative than positive information. Another study found that a pessimistic tone influences readers' decisions in a statistically meaningful way (Garcia, 2013).

Researchers used various methods to measure the tone of writing in managers' reports. 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 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 and Turney (2013) dictionary, contains the general word list (Bozorg Asl et al., 2021).

This study broadly refers to the information literature of the Annual Reports (Brown and Taker, 2011; Cole and 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. 2013). 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 study's most important finding 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 ongoing concern problems.

Jo and Shin (2016) suggest 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 quantitative data such as financial ratios and qualitative data 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 incorporation 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 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 enhances 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 during 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 should 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 (2017) 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.

Elsayed and Elshandidy (2020) explored the incremental explanatory power of narrative-related disclosures in predicting corporate failure by creating a comprehensive corporate failure-related lexicon. 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, precipitating 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 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, including 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 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.

Visvanathan's (2021) study aligns 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 using qualitative and quantitative criteria to predict bankruptcy. They concluded that qualitative information from 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, as well as the tone of the auditors' report, have been investigated. Still, those investigations have not been carried out simultaneously or in Iran's economic environment 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.

2.3 Research hypotheses

The two hypotheses of the present study are:

1. Using the text mining approach, the ability to predict the going concern of companies with/without growth opportunities is greater than the data mining approach.
2. The ability to predict companies with/without growth opportunities using the text mining approach significantly differs from expecting the going concern using the data mining approach.

3. Research Methodology

The research is applied in terms of purpose because its results can be used by potential and actual investors and other groups, and it is correlational 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 McDonald (2015) dictionary were then used to process the text analysis section. After extracting the research's numerical and textual data, EViews software version 10 was used to test the hypotheses.

3.1 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 using the systematic elimination method, 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.

3.2 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 McDonald dictionary (2015), available through the relevant site and contains 354 positive and 2355 negative words. The translation of this dictionary was used to analyze 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 (desirable, excellent, and profit) and negative words (unfavorable, weakness, and loss) can be counted in accounting narrations.

The frequency of positive and negative words reflects 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. Afterwards, 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 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.

3.3 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. Suppose the result is greater than 0.5 decimal places (used for the company's equal index of bankruptcy or non-bankruptcy). In that case, the company is less likely to continue as a going concern. Researchers who have used this method include Mayew et al. (2015) and Li and Wang (2017).

3.4 Vuong Z Test

For comparison of the power of two models in a common statistical sample. Considering the coefficient of determination obtained from estimating the two models is necessary. Because the amount of the coefficient of determination determines the model's power, a model with a higher determination coefficient has more power in explaining and forecasting the dependent variable.

However, a test must be carried out to determine whether the difference in the coefficients of determination of the two models is statistically significant. The desired test for comparing the difference between the coefficients of determination of both models was introduced by Vuong and is known as the Vuong Z test (Banimahd et al., 2016).

3.5 Research variables and models

3.5.1 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 due to losses so that the issue of the company's survival or liquidation can be discussed and voted on.

3.5.2 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, presented in Table (1) and 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), Li and Wang (2017), Rowland et al. (2021)*	*Reta*
net profit to total assets ratio	*Elsayed and Elshandidy (2020), Rowland et al. (2021)*	*Neta*
operating profit to total assets ratio	*Mayow et al. (2015), Li and Wang (2017), Rowland et al. (2021)*	*Ebitta*
current assets to current liabilities ratio	*Li and Wang (2017), Elsayed and Elshandidy (2020), Rowland et al. (2021)*	*Cacl*
working capital to total assets ratio	*Mayow et al. (2015), Rowland et al. (2021)*	*Wcta*
Total liabilities to total assets ratio	*Li and Wang (2017), Rowland et al. (2021)*	*Tlta*
sale revenue to total assets ratio	*Mayow et al. (2015), 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*

3.6 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)

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

The research model is based on research hypotheses and a text analysis approach, as shown in Equation 5.

Equation 5)

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

4. Analysis of research data and findings

4.1 Unit Root Test

Dummy regression occurs when nonstationary variables are present in the model. The test

presented by Levin et al. (2002) was used to evaluate the significance of the variables. When the time dimensions are large enough, this test is more efficient and powerful than other static tests (Najafzadeh et al., 2024). 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	-4.346	0.000
Neta	-7.215	0.000
Ebitta	-6.817	0.000
Cacl	-8.162	0.000
Wcta	-7.803	0.000
Tlta	-6.893	0.000
Saleta	-10.888	0.000
growth	-7.313	0.000
Posmda	-7.845	0.000
Negmda	-9.043	0.000
Gc_aud	-7.447	0.000

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). 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 linear relationships of the explanatory and dependent variables 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	-21.784	0.000
Error-values of text mining regression model	-24.212	0.000

4.2 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 the text mining approach has a coefficient of determination of 64 percent.

Table 4. Comparision analysis

Variables	Data mining model	Text mining model
C	*-0.260*	*-0.990*
Reta	5.310***	5.270***
Neta	-4.680	-6.380
Ebitta	0.660	2.240
Cacl	-3.160***	-3.290***
Wcta	2.340***	2.480***
Tlta	-0.760**	-0.740**
Saleta	-0.490***	-0.530***
Growth	0.630*	0.580*
Posmda		-0.250
Negmda		0.450*
Gc_aud		1.130***
R-squared	0.620	0.640

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

These coefficients indicate that 3 qualitative variables, positive words, negative words, and the auditor's expression of doubt, along with quantitative model variables, improve the model's explanatory power. Although the variable of positive words is not significant, 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.

4.3 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

Vuong Statistic Value	Z Statistics	Significance Level(prob.)	Test Result
-2.547	0.010	0.050	*The hypothesis is confirmed*

5. Conclusions and suggestions

Text mining is obtaining high-quality information from unstructured or semi-structured texts or data (Hearst, 2003). Accordingly, in 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., 2015; Huang 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 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 based on the results obtained and the corresponding coefficients. Furthermore, the findings of this study are consistent with those of Mayow et al. (2015), Elsayed et al. (2020), Lohmann and Ohliger (2020), Viswanathan (2021), and 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 when 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 ongoing 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, a set of conditions and cases 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 writing tone in financial research. Thus, due to the use of English dictionaries in translation and the linguistic differences, the reliability of research tools would increase if there were a standard dictionary in Persian.

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

3. According to all the text descriptions in the report, the research results 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:

1. Altman, E.I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), pp. 589-609. https://doi.org/10.2307/2978933
2. Altman, E.I., Sabato, G., and Wilson, N. (2010). The value of nonfinancial information in small and medium-sized enterprise risk management. *Journal of Credit Risk*, 2(1), pp. 95-127. https://api.semanticscholar.org/CorpusID:168632861
3. Banimahd, B., Arabi, M. and Hasanpor, Sh. (2016). Experimental research and methodology in accounting, Tehran, *Termeh Publications*. https://www.termehbook.com/product/9789649785547/

4. Beaver, W. H., McNichols, M. F., and Rhie, J. W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies*, 10(1), pp. 93-122. https://doi.org/10.2139/ssrn.634921
5. Bozorg Asl, M., marfo, M., and 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), pp. 79-107. https://doi.org/10.22054/qjma.2021.57363.2211
6. Brown, S. V. and J. W. Tucker. (2011). Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. *Journal of Accounting Research*. 49(2). pp. 309-346. https://doi.org/10.1111/j.1475-679X.2010.00396.x
7. Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63* (6), pp. 2899-2939. https://doi.org/10.1111/j.1540-6261.2008.01416.x
8. Carson, E., Fargher, N.L., Geiger, M.A., Lennox, C., Raghunandan, K., and Willekens, M. (2013). Audit Reporting for Going-Concern Uncertainty: A Research Synthesis. *Auditing-a Journal of Practice & Theory*, 32(1), pp. 353-384. https://doi.org/10.2308/AJPT-50324
9. Cole, C. J., and Jones, C. L. (2005). Management discussion and analysis: A review and implications for future research. *Journal of Accounting Literature*, 24(1), pp. 135-174. https://www.proquest.com/openview/ce689fdf754d0f04d7b8e555e6e5cbe0/1?pq-origsite=gscholar&cbl=31366
10. Davis, A. K., Ge, W., Matsumoto, D., and Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20(2), pp. 639–673. https://doi.org/10.1007/s11142-014-9309-4
11. Dey, L., Meisheri, H., and Verma, I. (2017). Predictive Analytics with Structured and Unstructured data - a Deep Learning based Approach. *IEEE. Informatics Bull.*, 18(1), pp. 27-34. https://api.semanticscholar.org/CorpusID:52012652
12. Elsayed, M., and Elshandidy, T (2020). Do narrative-related disclosures predict corporate failure? Evidence from UK nonfinancial publicly quoted firms. *International Review of Financial Analysis*, 71(1), A. 101555. https://doi.org/10.1016/j.irfa.2020.101555
13. Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, 15(4), pp. 915–953. https://doi.org/10.1007/s11142-009-9111-x
14. Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68(3), pp. 1267–1300. https://doi.org/10.1111/jofi.12027
15. Grunert, J.P., Norden, L. and Weber, M. (2005). The role of nonfinancial factors in internal credit ratings. *Journal of Banking and Finance*, 29(2), pp. 509-531. https://doi.org/10.1016/j.jbankfin.2004.05.017
16. Gutierrez, E., Krupa, J., Minutti-Meza, M., and Vulcheva, M. (2020). Do going concern opinions provide incremental information to predict corporate defaults?. *Review of Accounting Studies*, *25*(4), pp. 1344-1381. https://doi.org/10.1007/s11142-020-09544-x
17. Hamer, M.M. (1983). Failure prediction: sensitivity of classification accuracy to alternative statistical methods and variable sets. *Journal of Accounting and Public Policy*, 2(4), pp. 289-307. https://doi.org/10.1016/0278-4254(83)90032-7
18. Hearst, M. (2003). What is text mining. SIMS, UC Berkeley, 5. https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf
19. Henry, E. (2008). Are investors influenced by how earnings press releases are written?. *The Journal of Business Communication*, 45(4), pp. 363-407. https://doi.org/10.1177/0021943608319388
20. Huang, X., Teoh, S. H., and Zhang, Y. (2014). Tone management. *The Accounting Review*,

89(3), pp. 1083–1113. https://doi.org/10.2308/accr-50684
21. Jo, N., and Shin, K. (2016). Bankruptcy prediction modeling using qualitative information based on big data analytics. *Journal of Intelligence and Information Systems*, 22(2), pp. 33-56. https://doi.org/10.13088/jiis.2016.22.2.033
22. Kashanipoor, M., Aghaee, M.A., and Mohseni Namaghi, D. (2020). Information Disclosure Tone and Future Performance. *Accounting and Auditing Review*, 26(4), pp. 570-594. (In Persian). https://doi.org/10.22059/acctgrev.2020.278084.1008146
23. Kou, J. (2022). Analysing Housing Price in Australia with Data Science Methods *(Doctoral dissertation, Victoria University)*. https://vuir.vu.edu.au/id/eprint/43940
24. Kumar, L., and Bhatia, P. (2013). Text mining: concepts, process, and applications. *Journal of Global Research in Computer Sciences*, 4(3), pp. 36-39. https://api.semanticscholar.org/CorpusID:58813172
25. Law, K. K., and Mills, L. F. (2015). Taxes and financial constraints: Evidence from linguistic cues. *Journal of Accounting Research*, 53(4), pp. 777–819. https://doi.org/10.1111/1475-679X.12081
26. Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of accounting research*, 48(5), pp. 1049-1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x
27. Li, F. (2011). Textual Analysis of Corporate Disclosures: A Survey of the Literature. *Journal of Accounting Literature*, 29(1), pp. 143-165. http://www.cuhk.edu.hk/acy2/workshop/20110215FengLI/Paper1.pdf
28. Li, Y., and Wang, Y. (2017). Machine learning methods of bankruptcy prediction using accounting ratios. *Open Journal of Business and Management*, 6(1), pp. 1-20. https://doi.org/10.4236/ojbm.2018.61001
29. Leshno, M., and Spector, Y. (1996). Neural network prediction analysis: the bankruptcy case. *Neurocomputing*, 10(2), pp. 125-147. https://doi.org/10.1016/0925-2312(94)00060-3
30. Levin, A., Lin, C.F., and Chu, C-S.J. (2002). Unit root tests in panel data. *Journal of Econometrics*, 108(1), pp. 1-24. https://doi.org/10.1016/S0304-4076(01)00098-7
31. Lohmann, C., and 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), pp. 137-172. https://doi.org/10.1007/s11573-019-00938-1
32. Lopatta, K., Gloger, M. A., and Jaeschke, R. (2017). Can language predict bankruptcy? The explanatory power of tone in 10-K filings. *Accounting Perspectives*, 16(4), pp. 315-343. https://doi.org/10.1111/1911-3838.12150
33. Loughran, T., and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), pp. 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x
34. Loughran, T., and McDonald, B. (2015). The Use of Word Lists in Textual Analysis. Journal of Behavioral Finance, 16(1), pp. 1–11. https://doi.org/10.1080/15427560.2015.1000335
35. Mayew, W.J., Sethuraman, M., and Venkatachalam, M. (2015). MD&A Disclosure and the Firm's Ability to Continue as a Going Concern. *The Accounting Review*, 90(4), pp. 1621-1651. https://doi.org/10.2308/accr-50983
36. Mirali, M., Gholami Moghaddam, F., and Hesarzadeh, R. (2018). Investigation of the Relationship between Financial Reporting Tone with Future Corporate Performance and Market Return. *Financial Accounting Knowledge*, 5(3), pp. 81-98. (In Persian). https://doi.org/10.30479/jfak.2018.1513
37. Mohammad, S. M., and Turney, P. D. (2013). Crowdsourcing a word–emotion association

lexicon. *Computational intelligence*, 29(3), pp. 436-465. https://doi.org/10.1111/j.1467-8640.2012.00460.x
38. Namazi, M., Hajiha, Z., & Chenaribokat, H. (2018). Modeling and Identifying Hierarchy of the Effective Measures of the Earning Management on the Prediction of the Bankruptcy. *Financial Management Strategy*, 6(4), pp. 1-27. https://doi.org/10.22051/jfm.2018.13604.1257
39. Najafzadeh, A., Farzinvash, A. A., Yousefi sheikh robaat, M., and Elahi, N. (2024). The Mechanism of Fiscal Policy Transfer in the Economy: Evidence of the Asymmetric Behavior of the Fiscal Expenditure Multiplier During Business Cycles. *The Journal of Economic Studies and Policies*, 10(2), pp.157-187. https://doi.org/10.22096/esp.2024.535691.1555
40. Nießner, T., Gross, D. H., and 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. https://doi.org/10.3390/jrfm15100459
41. Odom, M.D., and Sharda, R. (1990). A neural network model for bankruptcy prediction. *Joint international conference on neural networks*, 2(1), pp. 163-168. https://doi.org/10.1109/IJCNN.1990.137710
42. Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 18(1), pp. 109-131. https://doi.org/10.2307/2490395
43. Pervan, I., and Kuvek, T. (2013). The relative importance of financial ratios and nonfinancial variables in predicting insolvency. *Croatian Operational Research Review*, 4(1), pp. 187-197. https://hrcak.srce.hr/file/143355
44. Rowland, Z., Kasych, A., and Suler, P. (2021). Prediction of financial distress: case of mining enterprises in Czech Republic. *Ekonomicko-manazerske spektrum*, *15*(1), pp. 1-14. https://doi.org/10.26552/ems.2021.1.1-14
45. Shaw, M.J., and Gentry, J.A. (1990). Inductive learning for risk classification. IEEE Expert, 5(1), pp. 47-53. https://doi.ieeecomputersociety.org/10.1109/64.50856
46. Shin, K., Lee, T.S., and Kim, H. (2005). An application of support vector machines in the bankruptcy prediction model. *Expert syst. Appl.*, 28(1), pp. 127-135. https://doi.org/10.1016/j.eswa.2004.08.009
47. Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1), pp.101-124. https://doi.org/10.2139/ssrn.171436
48. Siano, F., and Wysocki, P. (2021). Transfer learning and textual analysis of accounting disclosures: Applying big data methods to small (er) datasets. *Accounting Horizons*, 35(3), pp. 217-244. https://doi.org/10.2308/HORIZONS-19-161
49. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), pp. 1139-1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x
50. Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), pp.1437-1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x
51. 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), pp. 223-239. https://doi.org/10.1057/s41310-021-00107-3
52. Wang, C. J., Tsai, M. F., Liu, T., and Chang, C. T. (2013, October). Financial sentiment analysis for risk prediction. *In Proceedings of the Sixth International Joint Conference on Natural Language Processing*. pp. 802-808. http://aclweb.org/anthology/I/I13/I13-1097.pdf

53. Yekini, L. S., Wisniewski, T. P., and Millo, Y. (2016). Market reaction to the positiveness of annual report narratives. *The British Accounting Review*, 48(4), pp. 415–430. https://doi.org/10.1016/j.bar.2015.12.001
54. Zhang, Z., Luo, M., Hu, Z., and 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. https://doi.org/10.3390/app12136662
55. Zhao, Y., Wei, S., Guo, Y., Yang, Q., and Kou, G. (2022). FisrEbp: Enterprise Bankruptcy Prediction via Fusing its Intra-risk and Spillover-Risk. arXiv preprint arXiv, 2202. https://api.semanticscholar.org/CorpusID:246652179
56. Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, 22(1), pp. 59-82. https://doi.org/10.2307/2490859