

Predicting likelihood of fraud among financial distressed firms in Malaysia using textual analysis

Marziana Madah Marzuki¹, Syerina Azlin Md Nasir², Siti Fadilah Mat Zain^{1,3},
Nik Siti Madihah Nik Mangsor²

¹Faculty of Accountancy, Universiti Teknologi MARA Cawangan Kelantan, Machang, Malaysia

²College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Kelantan, Kota Bharu, Malaysia

³Accounting Department, CSD Solutions Sdn. Bhd., Kajang, Malaysia

Article Info

Article history:

Received Oct 18, 2023

Revised Dec 2, 2023

Accepted Dec 25, 2023

Keywords:

Annual report

Financial distressed firms

Fraud

Text clustering

Topic modeling

ABSTRACT

This research paper aims to analyze and predict fraud patterns among failed companies in Malaysia. The approach involves utilizing textual analysis on the management discussion and analysis (MD&A) section within the annual reports. The dataset is subjected to text clustering to group companies based on similar financial characteristics. This clustering process entails several steps, including data conversion, collation, and summarization into a structured format, followed by text pre-processing to cleanse the dataset. Notably, RapidMiner Studio software was utilized to extract data for the study. Subsequently, the documents are clustered using both the K-means and latent dirichlet allocation (LDA) methods. Upon examining a sample of 22 failed companies in the year 2020, the study reveals that financially distressed companies exhibit prominent financial negativity and utilize litigious financial terms within their MD&A sections. These linguistic traits are found to be closely associated with seven distinct characteristics of fraudulent firms. This preliminary findings provide compelling evidence that financial pressure may serve as a triggering factor for fraudulent activities within companies.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Syerina Azlin Md Nasir

College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Kelantan

Kota Bharu Campus, Lembah Sireh, 15050 Kota Bharu, Kelantan Darul Naim, Malaysia

Email: syerina@uitm.edu.my

1. INTRODUCTION

Fraud is a global problem which has been discussed in worldwide. The critical component of fraud is fraudulent financial reporting which needs to have effective fraud prevention and detection system. ACFE (2022) reported that from 2,110 cases investigated, the misappropriation of assets shows the highest cases reported with 86% of the cases followed by corruption with 50% of the cases and fraudulent financial reporting with 9% of the cases. However, the median losses occurred due to fraudulent financial reporting was very huge with \$593,000, followed by corruption median loss by \$150,000 and misappropriation of assets by \$100,000.

Various academic research has been done in predicting fraudulent financial reporting including the developed model which used financial ratios in predicting different events of fraud, earning manipulation, earning management and bankruptcy. The models include are by [1], Altman Z-score (1965) and the recent model is F-score developed by [2]. All these models used financial and non-financial information in predicting the likelihood of fraud. Goel and Uzuner [3] argued that detecting fraud is a complex problem and

not one of predictors will always be successful in fraud detection because once the fraud indicators are publicly known, the companies can find another and more creative ways to conceal the fraud.

Recently, there has been growing interest among researchers in exploring qualitative textual content in annual reports to detect fraud indicators [4]–[7]. This textual analysis method has been extensively studied in developed countries like the United States and China, focusing on specific sections of the annual report, particularly the management discussion and analysis (MD&A) section, to identify fraudulent practices [4], [6], [8]–[10]. However, such research remains underexplored in developing countries like Malaysia, particularly concerning the role of financial reporting in fraud detection.

Malaysia offers a unique institutional setting, influenced by both common-law and code-law countries, which affects reporting incentives. Despite adopting reporting standards from common-law countries, Malaysia shares similarities with code-law countries, leading to a lower level of enforcement due to institutional characteristics like family and political connections [11]–[13]. Jurisdiction-level institutional factors and firm-level factors significantly influence reporting incentives in Malaysia [14], [15]. However, only a limited number of studies have focused on textual analysis for fraud prediction in Malaysia. Non and Azis [16] examine whether Malaysian public listed companies expressed any specific sentiments in the management discussion and analysis section of the companies' annual report during the COVID-19 pandemic. The sentiments are extracted by means of computer-automated textual analysis through the linguistic inquiry and word counts and the Loughran–McDonald financial sentiment dictionary. Despite of using the textual analysis in the MD&A section of Malaysian annual report to analyze fraud, this study focuses on the sentiments that the companies adopted when communicating with their stakeholders. Therefore, this study fills the gap. The use of textual analysis in accounting and finance is currently in initial stage especially in the emerging market such as Malaysia. However, it has great potential to be applied due to significant volumes of documentation that are used to disclose economic and financial information such as financial statements, audit reports, corporate social reports, management reports, accounting standards or analyst's reports among others, which rely more on the value of textual, just nor numerical data [17]. Additionally, the use of digital tools and social media by companies has significantly increased the volume of unstructured documents that are available on the internet.

Textual analysis has gained significant traction in finance literature due to the inherent challenges in accurately quantifying qualitative information [18]. This method has become increasingly popular among practitioners for analyzing the information content of corporate disclosures, including conference calls, earnings press releases, and annual reports [9]. Textual sources may contain information that does not support the perception from the financial statements, either because there is an intention to "sweeten" the actual financial situation and performance of the company or because it is unintentionally exposing misleading quantitative information. This can be used to look for hidden cues in the corporate disclosures [19]. For example, the hidden cues related to differences between quantitative and textual information are the use of imagery, pleasantness or ambiguity in textual information [8].

Numerous studies have focused on the MD&A sections of annual reports to explore language-based methods for distinguishing between fraudulent and non-fraudulent firms and identifying failing and non-failing companies [5]–[7], [20]–[22]. For instance, [18] examined language usage differences between fraud firms and those self-reporting material weaknesses in internal controls. They employed two word classification methods: proportional weights and the term weighting procedure. While the proportional weights approach did not yield significant coefficients in the seven-word lists, the term weighting procedure revealed significant associations between fraud firms and specific word categories. Purda and Skillicorn [9] developed a language-based method using MD&A sections to detect fraud, achieving high correct classification rates with certain words identified as top predictors. Goel and Uzuner [3] investigated the significance of sentiment in MD&A for fraud detection, finding that fraud firms exhibited more subjectivity, adjectives, and adverbs compared to non-fraudulent firms. Dong *et al.* [10] utilized Systemic Functional Linguistics theory to explore seven information types in MD&A, achieving an average prediction accuracy of 82.36%. Despite the importance of textual analysis in fraud prediction, its application in emerging countries remains limited. Notably, a study in China focused on textual analysis to investigate financial reports of listed companies, primarily examining the likelihood of administrative punishments based on non-financial influence [23].

In a recent study, [22] investigated the potential of a hierarchical attention network (HAN), a state-of-the-art deep learning (DL) model, for advanced fraud detection. The HAN method aimed to capture the content and context of managerial comments in the MD&A section, achieving superior performance compared to extreme gradient boosting (XGB) and artificial neural networks (ANN) in distinguishing fraudulent and non-fraudulent firms. Meanwhile, Li *et al.* [7] conducted a study focusing on Chinese manufacturing firms and employed three language dimensions from MD&A sections, combined with financial indicators, to detect financial statement fraud. The study emphasized the significance of textual

quality, forward-looking information, and positive sentiment in improving fraud detection accuracy. Furthermore, the analysis of MD&A sections revealed that fraud companies with more significant similarities to previous MD&A sections were more likely to face administrative punishments. Conversely, fraud companies showing more significant similarities to previous non-MD&A sections had a lower probability of such penalties, highlighting the importance of the MD&A section in determining fraud characteristics.

This study addresses a gap in the existing literature by examining the use of textual information in emerging economies and contributing to the detection of financial fraud from both theoretical and empirical perspectives. By concentrating on companies experiencing financial difficulties, this research aids regulators in the early prediction of potential fraudulent cases. Consequently, the aim of this investigation was to apply text analysis of MD&A to study patterns of fraud among unsuccessful companies as well as predict the likelihood of such occurrences. The focus was on businesses listed under practice note 17/2005 issued by bursa Malaysia for financially distressed firms that may resort to manipulating financial statements due to pressure stemming from their financial status. It is noted that pressure plays a significant role as a trigger factor leading to fraudulent activities [24]. The research contributes to the existing body of knowledge on using textual analysis for predicting fraud, demonstrating its effectiveness in emerging economies such as Malaysia. Additionally, it expands the scope of analysis to anticipate the likelihood of fraud among vulnerable companies, providing valuable information for regulators to identify fraudulent patterns early.

Our research has made several contributions. It expanded the body of knowledge on using textual analysis to predict fraud. Additionally, it presented evidence indicating the potential for textual analysis to predict fraudulent activity in developing countries like Malaysia, which has been an overlooked area in previous studies that focused mainly on developed countries. Moreover, our study broadened the scope by predicting fraud among companies with fraud potential rather than actual cases of fraud. This emphasis offered regulators valuable insights into identifying patterns or characteristics of companies at risk of perpetrating fraud before such instances are officially reported. The paper is structured as follows: section 2 provides an overview of the literature related to fraud and the significance of textual data in financial reporting. Section 3 outlines the research methodology, while section 4 presents the findings. Lastly, section 5 delves into the implications of this research.

2. METHOD

The proposed method consists of a few steps such as, sample selection, dataset collection, data analysis, text pre-processing and last text analysis which consist of document clustering and topic modelling techniques respectively. RapidMiner was utilized for this experiment due to its prominence as a data mining software. This method is further detailed in the following subsections.

2.1. Sample selection, data collection and data analysis

This study utilized a dataset from 22 PN17 companies in Malaysia. Utilizing established algorithms from the data mining software RapidMiner Studio, the annual reports were analyzed to ensure impartiality and concentrate exclusively on 2020 data. The MD&A text datasets from each company's 2020 annual report were gathered, structured, and cleaned for analysis purposes. Table 1 outlines the PN17 companies included in this research for the specified year.

Table 1. List of PN17 companies in Malaysia

PN17 companies					
1	EKA Noddles	9	Nationwide Express	17	Lotus KFM
2	TH Heavy	10	Scomi Group	18	IDIMENSION Berhad
3	Asia Media	11	Scomi Energy	19	Dolomite Corp
4	Barakah Offshore	12	Perak Corporation	20	MAA Group
5	Bertam Alliance	13	Iqzan Holding	21	APFT
6	Daya Material	14	FSBM Holding	22	G NEEPTUNE
7	Consortium Transnasional	15	Comintel Corporation		
8	Malaysia Pacific Corporation	16	Brahim Holding		

2.2. Text preprocessing

Text preparation plays a crucial role in text mining as it aims to standardize various forms of words into one cohesive form. Moreover, techniques for text preprocessing carry substantial importance and have been extensively researched in the field of machine learning to create a refined corpus. The cleaning process was carried out to uphold the coherence of the corpus documents through the removal of duplicates and less

informative phrases. Furthermore, this procedure aimed at rectifying typographical errors and validating or correcting values based on a predefined list of entities. Preprocessing comprises several steps that are undertaken to prepare the textual data for qualitative analysis, including:

- i) Normalization: the data should be normalized or standardized to bring all of the variables into proportion with one another. Non-numeric qualitative data should be converted to numeric quantitative data.
- ii) Tokenization: removes extra white space and breaks text apart to identify meaningful units, called tokens-usually words but can also be phrases (n-grams) and emoticons.
- iii) Stopwords removal: removes extremely common words that are considered to be unlikely to provide value for document retrieval or analysis.
- iv) Stemming: reduces vocabulary size by using heuristic algorithms to “remove morphological affixes from words, leaving only the word stem”.
- v) Building corpus: each document is represented in the corpus by a sequence of pairings. The first digit of the pair conveys that the numeric ID relates to a word, while the second digit expresses how frequently that word occurs.

The text data in RapidMiner Studio underwent several pre-processing steps to prepare for qualitative analysis. These operations, shown in Figure 1, included tokenize, transform cases, filter tokens, filter stopwords (English), and stem (porter). Tokenization was utilized to create individual tokens consisting of single words. The transform cases operator converted all characters to lowercase for consistency. Additionally, the filter tokens (by length) function filtered token sizes by excluding those that did not meet specified criteria. In addition, the operator filter stopwords (English) was used to eliminate English stopwords from the documents, and specified regular expressions were substituted using the replace tokens operator. The last two sub-processes involved extracting length and token number of each document through the extract length and extract token number operators, respectively. This metadata will be utilized in subsequent analysis stages.

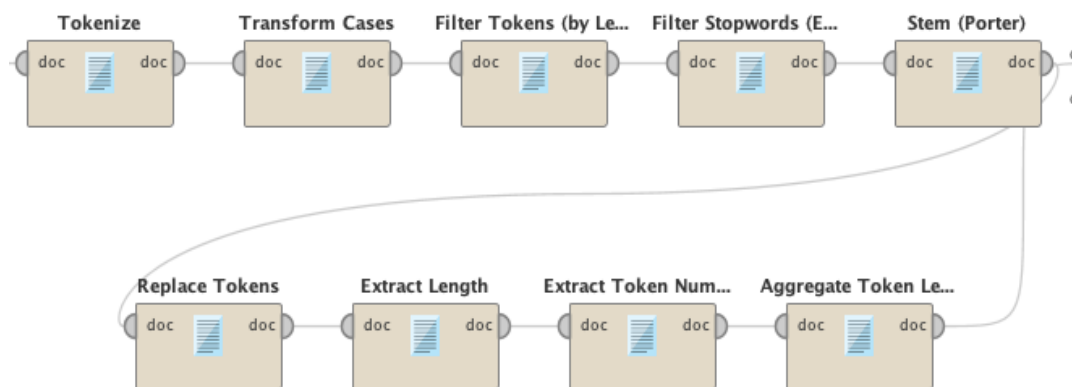


Figure 1. Operators used in text pre-processing

2.3. Text analysis

In this study, we utilized a combined method for text analysis by integrating document clustering and topic modeling to improve results. Recent studies by [25], [26] indicate that this integration can boost the performance of clustering. As mentioned in [27], document clustering and topic modeling complement each other, with the former producing unlabeled clusters and the latter effectively characterizing and explaining these clusters. Improved clustering outcomes lead to more informative and coherent topic models, thus contributing to enhanced overall results.

The document clustering process involves the grouping of similar documents together. We employed unsupervised machine learning, specifically the k-means algorithm, on unlabelled data to achieve this. The k-means algorithm automatically forms k clusters within the corpus by assigning each document to the most similar cluster based on cosine similarity measures [28]. Cosine similarity is a preferred measure for text analysis, as it quantifies document similarity based on the distance between documents. The cosine similarity is calculated based on the angle between two vectors, reflecting their direction. The cluster centroids are recalculated by averaging all documents within each cluster, until they reach maximum optimization or remain unchanged. Figure 2 demonstrates the sequential procedures involved in document clustering with RapidMiner, which enables the efficient incorporation of these methods for textual analysis.

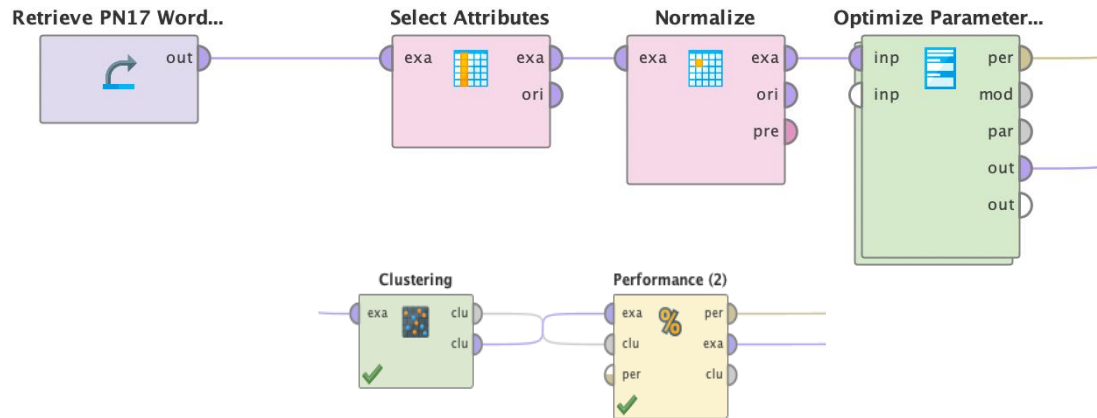


Figure 2. Document clustering process in RapidMiner

Once the document clustering is complete, it is found that the clusters generated lack descriptive information or characterization. Hence, topic modeling is applied to further analyze and interpret the clusters. Topic modeling is a powerful technique that can help uncover hidden themes and patterns within a set of documents. This technique assists in unveiling the underlying structure of a group of documents by identifying a range of topics, where each topic is depicted as a distribution over a set of words [25]. Among different methods for topic modeling, latent dirichlet allocation (LDA) holds widespread popularity in text mining owing to its exceptional adaptability and generalization capabilities [29]. The concept behind LDA involves representing a textual document as an amalgamation of multiple topics, each containing various words. For LDA's functionality, it necessitates input text documents and the expected number of topics.

The study utilizes LDA with Gibbs Sampling, which is suitable for situations where the joint distribution is unknown or challenging to sample directly [30]. The LDA method randomly assigns each word in the 22 documents to a topic. Subsequently, it computes distributions of topics across documents and words within topics based on word frequencies across all topics and the frequency of each topic in the document. The prevalence of each topic within the document is determined through frequency counts during initialization (topic frequency), as well as considering the dirichlet-generated multinomial distribution over topics for each document. The occurrence of each word in each topic is evaluated based on the frequency counts determined at the beginning (word frequency) and a dirichlet-generated multinomial distribution over words for each topic. The research emphasizes calculations regarding words and topics, wherein the assignment of each word to a topic is re-assigned iteratively according to the most substantial conditional probability. This iterative process continues until stable assignments of words to topics are achieved. In RapidMiner, this procedure is carried out using the extract topic from data operator, as depicted in Figure 3.

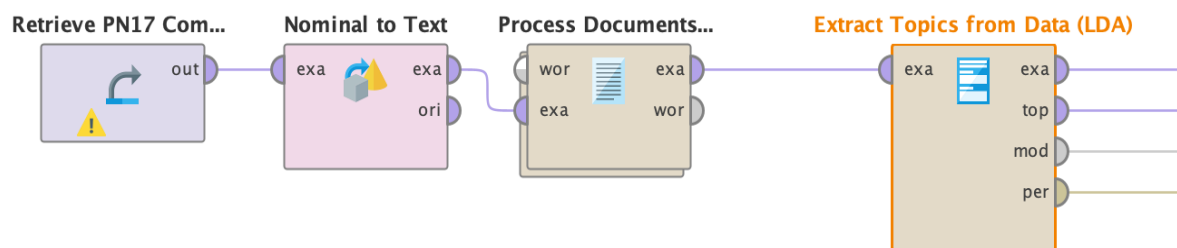


Figure 3. Topic modeling process in RapidMiner

3. RESULTS AND DISCUSSION

3.1. Result of text pre-processing

The section discusses the results of the data cleansing process carried out in the research, where three important numerical values are used during text processing. These values consist of document length, token number, and token length, which will play a vital role in subsequent text clustering analysis. To

accomplish this task, RapidMiner software is utilized along with operators such as extract length, extract token number, and aggregate token length. Figure 4 provides a visual representation of the dataset before (Figure 4(a)) and after undergoing text cleansing (Figure 4(b)). As a result, a set of word vectors for PN17 companies is obtained to enable further analysis and exploration. All symbols employed in the equations will be defined in the following sections.

Row No.	COMPANY	YEAR	TEXT
1	EKA NOODLES	2020	For the FY 2020, the Group registered a total revenue of RM41.22 million, registering a significant incre...
2	TH HEAVY	2020	In 2020, the Group registered a revenue of RM61.9 million, an increase of RM4.4 million as compared t...
3	ASIA MEDIA	2020	The Group reported a Revenue level of RM0.4 million in financial period 1 October 2019 to 31 March 2...
4	BARAKAH OFFS...	2020	The Group has returned to the black with profit after tax of RM25.01 million for FY2020 (from 1 July 20...
5	BERTAM ALLIA...	2020	The Group registered a revenue of RM6.4 million for the financial year ended 31 December 2020. Year...
6	DAYA MATERIAL	2020	The Group's financial year end has been changed from 31 December to 30 June on 29 May 2019. As s...
7	CONSORTIUM ...	2020	For the financial year ended 31st December 2020, the Group recorded a revenue of RM30.4 million a ...
8	MALAYSIA PACI...	2020	The Group main activities are investment property and property development. For the financial year end...
9	NATIONWIDE E...	2020	For the financial year ended 31 March 2020, the Group's revenue declined whilst the Group's loss after...
10	SCOMI GROUP	2020	The comparative financial results for the financial year ended 30 June 2019 ("FY2019") is for a period o...
11	SCOMI ENERGY	2020	The current financial year ("FY20") represents 12 months while the comparative period comprises 15 m...
12	PERAK CORPO...	2020	Perak Corporation Berhad ("PERAK CORP" or "Company") is an investment holding company with the sub...
13	IQZAN HOLDING	2020	In the financial period under review, the Company has chosen to downsize in the packaging business an...
14	FSBM HOLDING	2020	FSBM Group reports a revenue of RM128,000 and loss before tax of RM665,000 for the financial year e...
15	COMINTEL CO...	2020	Revenue from continuing operations of the Group for FYE 2020 was RM1.9 million. The revenue has dec...

(a)

Row No.	text	COMPANY	YEAR	document_...	token_num...
1	group regist...	EKA NOODLES	2020	1974	349
2	group regist...	TH HEAVY	2020	488	85
3	group repor...	ASIA MEDIA	2020	990	169
4	group retur...	BARAKAH O...	2020	1520	261
5	group regist...	BERTAM ALL...	2020	333	60
6	group financ...	DAYA MATE...	2020	1748	302
7	financi year ...	CONSORTIU...	2020	364	66
8	group main ...	MALAYSIA P...	2020	513	92
9	financi year ...	NATIONWID...	2020	569	100
10	compar fina...	SCOMI GROUP	2020	1775	297
11	current fina...	SCOMI ENER...	2020	4084	709
12	perak corpo...	PERAK COR...	2020	952	168
13	financi perio...	IQZAN HOL...	2020	1102	186
14	fsbm group ...	FSBM HOLDI...	2020	307	56
15	revenu conti...	COMINTEL C...	2020	1354	231

(b)

Figure 4. Datasets (a) before cleaning process and (b) after cleaning process

3.2. Result of K-means clustering

Following the text preprocessing stage, a comprehensive collection of 799 word vectors was created and subjected to clustering using the K-means algorithm. In this clustering process, numerical values such as document size, token length, and token count were essential in computing distances between the word

vectors. To improve the effectiveness of the grouping or clustering process, the resulting outcomes were then normalized to decrease variable magnitude. This normalization step improves accuracy and efficiency in cluster analysis, as shown in Figure 5. Figure 5(a) depicts the dataset before the normalization process, while Figure 5(b) displays the dataset after the normalization process.

Row No.	COMPANY	text	document_...	token_num...	token_leng...
1	EKA NOODLES	group group...	6259	697	8.980
2	TH HEAVY	group group...	1536	169	9.089
3	ASIA MEDIA	group group...	3126	337	9.276
4	BARAKAH O...	group group...	4811	521	9.234
5	BERTAM ALL...	group group...	1047	119	8.798
6	DAYA MATE...	group group...	5534	603	9.177
7	CONSORTIU...	financi finan...	1146	131	8.748
8	MALAYSIA P...	group group...	1619	183	8.847
9	NATIONWID...	financi finan...	1793	199	9.010

(a)

Row No.	COMPANY	text	document_...	token_num...	token_leng...
1	EKA NOODLES	group group...	0.527	0.548	-0.366
2	TH HEAVY	group group...	-0.787	-0.784	0.083
3	ASIA MEDIA	group group...	-0.345	-0.360	0.855
4	BARAKAH O...	group group...	0.124	0.104	0.683
5	BERTAM ALL...	group group...	-0.923	-0.910	-1.116
6	DAYA MATE...	group group...	0.325	0.311	0.449
7	CONSORTIU...	financi finan...	-0.896	-0.880	-1.323
8	MALAYSIA P...	group group...	-0.764	-0.749	-0.915
9	NATIONWID...	financi finan...	-0.716	-0.708	-0.242

(b)

Figure 5. Dataset (a) before normalisation process and (b) after normalisation process

The next step involves determining the best value for K, involving clustering and performance evaluation using the optimize parameters operator. In this study, we utilized the K-means clustering technique to enhance proximity within clusters while maximizing their distance from each other. Analyzing textual data poses challenges due to its high-dimensional nature, making it difficult to determine the most suitable number of clusters. To address this issue, we adopted the "elbow method" for cluster selection, enabling us to assess clustering performance for different K values. The results indicated that a value of K=7 was the most favorable choice, as shown in Figure 6.

3.3. Result of LDA approach

Previous studies have predominantly concentrated on analyzing textual data to differentiate between fraudulent and non-fraudulent firms. However, there is a lack of research examining the text patterns of financially troubled companies, which are associated with an increased likelihood of fraud. It is essential for regulators to comprehend these textual patterns in order to anticipate potential fraud among such companies before they are formally classified as fraudulent entities. Moreover, the current methods for detecting fraud are often difficult to ascertain, underscoring the significance of undertaking this investigation.

The categories identified through the LDA method, specifically the gibbs sampling technique, illuminate the themes revealed in the text data. As previously mentioned, using the LDA approach necessitates text documents and a predetermined number of subjects. In this research, based on K-means

clustering results, seven was established as the optimal number of topics. The likelihood of each word within a "topic" is then organized in descending order according to their frequencies. The results from LDA offer valuable insights, particularly concerning the financial attributes of PN17 companies. Each document is assumed to be produced by multiple topics out of all possible ones that exist and signifies that every word in a document can be attributed to one of its topics.

Table 2 illustrates the successful application of topic modeling to cluster various topics from the chosen text documents of PN17 companies. The method employs latent dirichlet allocation with gibbs sampling to create a term×document matrix and generate the distribution of topics across words. Throughout this process, LDA assigns each word in the document to one of t themes for each text d until stability is reached, indicating a significant state. Domain expertise or tacit knowledge is essential for assigning meaningful topic labels by identifying common words within each topic.

The top ten most frequent words for each topic are presented in Table 2, offering insights into the associated concepts of each "topic". Topic 0 is related to performing an activity, while topic 1 is connected to loss and impairment activity. Furthermore, topic 2 concerns cost and acquisition activity, topic 3 relates to non-performing activity, and topic 4 is tied to business acquisition activity. Additionally, topic 5 is linked with income and profit activity, while topic 6 focuses on litigation activity. It's worth noting that out of the seven identified topics, five topics (topic 1, 2, 3, 4, and 6) are directly associated with negative financial aspects and litigation activities.

The five groups identified in the study correspond to the categories proposed by [18], namely loss and impairment (topic 1), cost and acquisition (topic 2), non-performing (topic 3), business acquisition (topic 4), and litigation (topic 6). It is important to note that the words in each cluster are closely related to the top-ranked distinguishing words for truthful and fraudulent reports as previously indicated [9], as shown in Table 2 (bold and italicized). Some of these words encompass acquisitions, sales, revenues, shares, increased, decreased, tax profit, and activities. The results suggest that while PN17 companies are not officially classified as fraudulent entities; the language used in their MD&A sections reflects terms commonly linked with fraud anticipation.

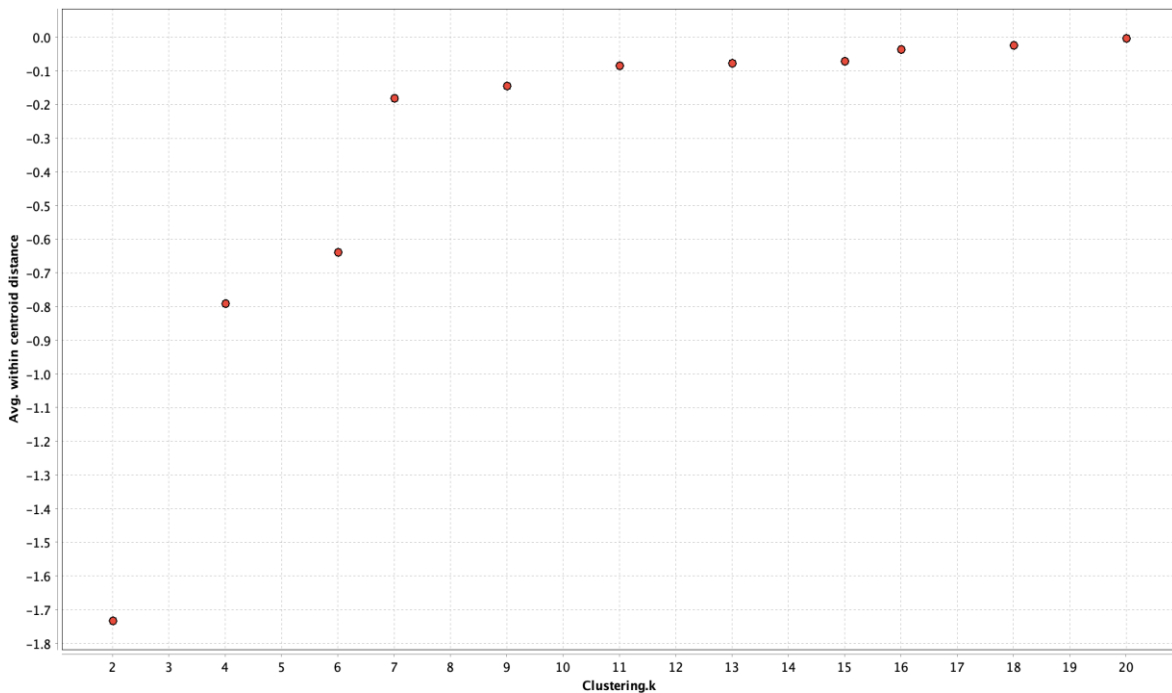


Figure 6. Scatter plot of K value

Based on the findings, Table 3 shows the specific financial activities linked to each PN17 company, revealing the distribution of topics within their documents. This examination provides a deeper understanding of the financial traits displayed by each company in 2020. For example, EKA Noodles falls under the income and profit activity category, with its subsidiaries allegedly involved in falsifying and modifying business records during 2020. These actions aimed to create uncertainty about the ownership of their product brand.

One of the major shareholders of the organization was alleged to be involved in illegal actions that negatively impacted its operations. These activities included directing the falsification and alteration of company documents, causing confusion among both the public and the market [31]. This suggests that there is a possibility that positive financial language may be used by the company to hide illicit income-generating activities with an aim to attract more investment.

Table 2. Top ten most frequent words for each topic identified

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Performing an activity	Loss and impairment	Cost and acquisition	Non-performing activity	Business acquisition	Income and profit activity	Litigation activity
million	group	group	revenue	company	million	financial
compared	year	recorded	due	lho	net	company
current	loss	million	financial	proposed	total	capital
fye	impairment	December	mainly	bursa	profit	period
activities	assets	business	period	shares	equity	subsidiaries
increase	compared	review	tax	securities	mainly	management
cash	previous	taxation	lower	exercise	operating	plan
increased	financial	insurance	operations	berhad	value	group
sales	share	acquisition	sesb	main	investment	basis
revenue	decrease	performance	cost	malaysia	higher	court

On the contrary, Scomi Group and Scomi Energy fall under the non-performing and litigation activity. This is further confirmed when an investigation into a loan of RM64.33 million from Scomi Energy Services Bhd to its parent company, Scomi Group Bhd without board approval resulted in financial distress for SESB when the amount was not repaid by Scomi Group. The advances were found to be arranged through a third-party firm of independent auditors without Audit Committee approval, leading to a legal case [32].

Table 3. Categories of financial characteristics by each company for 2020

Company	Financial characteristics
EKA Noodles	5
TH Heavy	1
Asia Media	6
Barakah Offshore	0
Bertam Alliance	1
Daya Material	3
Consortium Transnasional	1
Malaysia Pacific Corporation	1
Nationwide Express	1
Scomi Group	3
Scomi Energy	6
Perak Corporation	1
Iqzan Holding	2
FSBM Holding	3
Comintel Corporation	1
Brahim Holding	1
Lotus KFM	0
IDIMENSION Berhad	4
Dolomite Corp	4
MAA Group	5
APFT	6
G NEEPTUNE	4

The data in Table 3 indicates that out of the twenty-two PN17 companies in 2020, only four were associated with positive financial language (Topic 0 and Topic 5) as suggested by [18]. These companies include EKA Noodles, Barakah Offshore, Lotus KFM, and MAA Group. Conversely, most of the companies were linked to negative financial terms. Specifically, three companies: Asia Media, Scomi Energy, and APFT were connected with Topic 6 concerning litigious financial activities. This supports the findings of [4], where fraudulent firms showed a notably higher frequency of negative words compared to non-fraudulent ones. Furthermore, this aligns with the conclusions drawn by emphasizing the strong association between adverse financial terminology and litigation-related terms among fraudulent firms as well as those disclosing weaknesses in their internal accounting controls. Therefore, it is evident that there is a strong correlation

between the use of negative financial language and the likelihood of fraudulent financial reporting among financially distressed companies in Malaysia.

4. CONCLUSION

In conclusion, this study delved into the analysis of MD&A through textual examination to uncover patterns within PN17 companies and anticipate potential fraudulent activities. By harnessing advanced computer technology and artificial intelligence, an extensive array of texts and research pertaining to accounting and financial matters were scrutinized. The results uncovered seven distinctive clusters that characterize the financial features of the 22 PN17 companies studied. Moreover, by identifying the top 10 words in each cluster, significant terms used in the MD&A reports concerning predictive language for fraud were successfully identified. These findings not only lend support to but also offer empirical evidence for the fraud diamond theory by demonstrating how financial pressure can serve as a crucial trigger for instances of fraudulent practices within corporate entities.

Our contribution to the literature involves a thorough exploration of identifying PN17 companies based on similar financial characteristics, serving as an effective predictor of fraud. This extends the existing body of work in textual analysis and fraud detection, providing valuable insights for practitioners to make informed choices regarding the implementation of fraud detection methods. Moreover, our findings can assist auditors in assessing fraud risk by highlighting that a high frequency of identified linguistic markers in annual reports may indicate an elevated probability of fraudulent activities within organizations. These insights are crucial in alerting auditors to companies at higher risk for potential fraudulent behavior, aiding them in mitigating risks effectively through targeted interventions and vigilant oversight.

Furthermore, our findings have practical implications for regulators, such as the securities exchange commission (SEC). By utilizing linguistic markers to assess the probability of fraud among companies and flagging potentially fraudulent firms displaying these markers, regulators can initiate further investigation. However, it's important to acknowledge a limitation of this study, as only one year of data was analyzed. Future research should consider extending the study over multiple years to gain a more comprehensive understanding of fraud patterns among each distressed firm and strengthen its conclusions.




REFERENCES

- [1] M. D. Beneish, "The detection of earnings manipulation," *Financial Analysts Journal*, vol. 55, no. 5, pp. 24–36, Sep. 1999, doi: 10.2469/faj.v55.n5.2296.
- [2] P. M. Dechow, W. Ge, C. R. Larson, and R. G. Sloan, "Predicting material accounting misstatements," *Contemporary Accounting Research*, vol. 28, no. 1, pp. 17–82, Jan. 2011, doi: 10.1111/j.1911-3846.2010.01041.x.
- [3] S. Goel and O. Uzuner, "Do sentiments matter in fraud detection? estimating semantic orientation of annual reports," *Intelligent Systems in Accounting, Finance and Management*, vol. 23, no. 3, pp. 215–239, May 2016, doi: 10.1002/isaf.1392.
- [4] X. T. T. Le and G. Teal, "A review of the development in defining corporate social responsibility," *Science and Technology Development Journal*, vol. 14, no. 2, pp. 106–115, Jun. 2011, doi:10.32508/stdj.v14i2.1935.
- [5] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decision Support Systems*, vol. 50, no. 3, pp. 585–594, Feb. 2011, doi: 10.1016/j.dss.2010.08.009.
- [6] Y. Zhang, A. Hu, J. Wang, and Y. Zhang, "Detection of fraud statement based on word vector: evidence from financial companies in China," *Finance Research Letters*, p. 102477, Sep. 2021, doi: 10.1016/j.frl.2021.102477.
- [7] J. Li, N. Li, T. Xia, and J. Guo, "Textual analysis and detection of financial fraud: evidence from Chinese manufacturing firms," *Economic Modelling*, vol. 126, pp. 106428–106428, Sep. 2023, doi:10.1016/j.econmod.2023.106428.
- [8] D. B. Skillicorn and L. Purda, "Detecting fraud in financial reports," in *2012 European Intelligence and Security Informatics Conference*, 2012, pp. 7–13.
- [9] L. Purda and D. Skillicorn, "Accounting variables, deception, and a bag of words: assessing the tools of fraud detection," *Contemporary Accounting Research*, vol. 32, no. 3, pp. 1193–1223, Oct. 2014, doi: 10.1111/1911-3846.12089.
- [10] W. Dong, S. Liao, and L. Liang, "Financial statement fraud detection using text mining: A Systemic Functional Linguistics theory perspective," in *Pacific Asia Conference on Information Systems, PACIS 2016 - Proceedings*, 2016, p. 11.
- [11] E. L. Black and J. J. White, "An international comparison of income statement and balance sheet information: Germany, Japan and the US," *European Accounting Review*, vol. 12, no. 1, pp. 29–46, May 2003, doi: 10.1080/096381802200001127.
- [12] J. M. G. Lara and A. Mora, "Balance sheet versus earnings conservatism in Europe," *European Accounting Review*, vol. 13, no. 2, pp. 261–292, Jul. 2004, doi:10.1080/0963818042000203347.
- [13] J. Gassen, R. U. Fülbier, and T. Sellhorn, "International differences in conditional conservatism – the role of unconditional conservatism and income smoothing," *European Accounting Review*, vol. 15, no. 4, pp. 527–564, Dec. 2006, doi: 10.1080/09638180601102107.
- [14] U. Brüggemann, J.-M. Hitz, and T. Sellhorn, "Intended and unintended consequences of mandatory IFRS adoption: a review of extant evidence and suggestions for future research," *European Accounting Review*, vol. 22, no. 1, pp. 1–37, May 2013, doi: 10.1080/09638180.2012.718487.
- [15] P. Brown, J. Preiato, and A. Tarca, "Measuring country differences in enforcement of accounting standards: an audit and enforcement proxy," *Journal of Business Finance & Accounting*, vol. 41, no. 1–2, pp. 1–52, Jan. 2014, doi: 10.1111/jbfa.12066.




- [16] N. Non and N. A. Aziz, "An exploratory study that uses textual analysis to examine the financial reporting sentiments during the COVID-19 pandemic," *Journal of Financial Reporting and Accounting*, vol. 21, no. 4, pp. 895–915, Mar. 2023, doi: 10.1108/jfra-10-2022-0364.
- [17] F. Amani and A. Fadlalla, "Data mining applications in accounting: a review of the literature and organizing framework," *International Journal of Accounting Information Systems*, vol. 24, pp. 32–58, Feb. 2017, doi: 10.1016/j.accinf.2016.12.004.
- [18] T. Loughran and B. McDonald, "When Is a liability not a liability? textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, Jan. 2011.
- [19] J. L. G. Cabedo and D. Huguet, "Combined effects of auditing and discretionary accruals on the cost of debt: evidence from Spanish SMEs," *SAGE Open*, vol. 11, no. 4, p. 215824402110525, Oct. 2021, doi: 10.1177/21582440211052558.
- [20] R. Feldman, S. Govindaraj, J. Livnat, and B. Segal, "Management's tone change, post earnings announcement drift and accruals," *Review of Accounting Studies*, vol. 15, no. 4, pp. 915–953, Aug. 2009, doi: 10.1007/s11142-009-9111-x.
- [21] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: using financial text as a predictor of financial events," *Decision Support Systems*, vol. 50, no. 1, pp. 164–175, Dec. 2010, doi:10.1016/j.dss.2010.07.012.
- [22] P. Craja, A. Kim, and S. Lessmann, "Deep learning for detecting financial statement fraud," *Decision Support Systems*, p. 113421, Oct. 2020, doi: 10.1016/j.dss.2020.113421.
- [23] A. Qian and D. Zhu, "Financial report similarity and the likelihood of administrative punishment: based on the empirical evidence of textual analysis," *China Journal of Accounting Studies*, pp. 1–23, Sep. 2019, doi: 10.1080/21697213.2019.1642604.
- [24] S. Y. Huang, C.-C. Lin, A.-A. Chiu, and D. C. Yen, "Fraud detection using fraud triangle risk factors," *Information Systems Frontiers*, vol. 19, no. 6, pp. 1343–1356, Apr. 2016, doi: 10.1007/s10796-016-9647-9.
- [25] R. Sabbagh and F. Ameri, "A framework based on k-means clustering and topic modeling for analyzing unstructured manufacturing capability data," *Journal of Computing and Information Science in Engineering*, vol. 20, no. 1, Sep. 2019, doi: 10.1115/1.4044506.
- [26] M. Alhawarat and M. O. Hegazi, "Revisiting K-Means and topic modeling, a comparison study to cluster Arabic documents," *IEEE Access*, vol. 6, pp. 42740–42749, Jan. 2018, doi: 10.1109/access.2018.2852648.
- [27] P. Y. Shotorbani, F. Ameri, B. Kulvatunyou, and N. Ivezic, "A hybrid method for manufacturing text mining based on document clustering and topic modeling techniques," in *IFIP advances in information and communication technology*, 2016, pp. 777–786, doi: 10.1007/978-3-319-51133-7_91.
- [28] K. Zainal, N. F. Sulaiman, and M. Z. Jali, "An analysis of various algorithms for text spam classification and clustering using RapidMiner and Weka," *International Journal of Computer Science and Information Security*, vol. 13, no. 3, p. 66, 2015.
- [29] F. Gürçan and N. E. Çağiltay, "Big data software engineering: analysis of knowledge domains and skill sets using LDA-based topic modeling," *IEEE Access*, vol. 7, pp. 82541–82552, Jan. 2019, doi: 10.1109/access.2019.2924075.
- [30] S. He, H. Shin, and A. Tsourdos, "Distributed multiple model joint probabilistic data association with gibbs sampling-aided implementation," *Information Fusion*, vol. 64, pp. 20–31, Dec. 2020, doi: 10.1016/j.inffus.2020.04.007.
- [31] "Eka Noodles units sued for alleges involvement in seeking to confuse public over brand ownership," *The Edge Markets*, 2020. [Online]. Available: <https://www.theedgemarkets.com/article/eka-noodles-sued-allegedly-faking-ownership-noodles-brand>
- [32] "Scomi Energy investigates the RM64.33 million advance to parent Scomi Group, and cuts ties with it," *The Edge Markets*, 2020. [Online]. Available: <https://www.theedgemarkets.com/article/scomi-energy-investigates-rm6433-mil-advance-parent-scomi-group-and-cuts-tie-it>.

BIOGRAPHIES OF AUTHORS






Marziana Madah Marzuki    is an Associate Professor at Faculty of Accountancy, Universiti Teknologi MARA (UiTM) Machang, Kelantan. She holds Ph.D. Degree in Financial Reporting and Corporate Governance. Her research areas are financial reporting, corporate governance, auditing, risk management and Islamic accounting. She has published in *Journal of Contemporary Accounting and Economics*, *Pacific Accounting Review*, *Social Responsibility Journal*, *Accounting Research Journals*, *Asian Review of Accounting*, *Journal of Islamic and Accounting Business Research (JIABR)*. She can be contacted at email: marzianamadah@uitm.edu.my.






Dr. Syerina Azlin Md Nasir    earned her Ph.D. in Information Technology from Universiti Teknologi MARA (UiTM), Malaysia, and completed her undergraduate studies at the University of Salford in the United Kingdom. She holds the position of Associate Professor within the College of Computing, Informatics, and Mathematics at UiTM Cawangan Kelantan, Malaysia, where she has served as a faculty member since 2004. She has been actively involved in research endeavors, including participation in conferences and workshops, as well as membership in the Business Data Analytics Research Group and IEEE. She is a certified trainer in data analyst from RapidMiner and data integration from Talend. Her previous research publications have centered on topics related to database technology, ontology construction, and mapping. She maintains a strong commitment to research, consultancy, and scholarly publications. Her principal areas of interest encompass data mining, data analytics, text and web mining. She can be contacted at email: syerina@uitm.edu.my.



Siti Fadilah Binti Mat Zin    hold Master of Accountancy from Universiti Teknologi MARA (UiTM) Cawangan Puncak Alam, Malaysia. She currently pursues her postgraduate studies in Doctor of Philosophy (Ph.D.) at UiTM Shah Alam, Selangor. Her research area is the fraud detection and prevention among Public Listed Companies in Malaysia. She is an Account Executive of a Small and Medium Enterprise (SME) Company in Selangor. Her Company interest in Gas Piping and Liquefied Petroleum Gas (LPG) industry in Malaysia. Previously she was an auditor in an Audit Firm at Kota Bharu, Kelantan for three years. She can be contacted at email: sfadilah1992@gmail.com.



Nik Siti Madihah Nik Mangsor    hold Bachelor of Science (Hons.) Computational Mathematics from Universiti Teknologi MARA (UiTM) Cawangan Kuala Terengganu, Malaysia. Currently she pursues her undergraduate studies in Master of Science (Computer Science) by research at UiTM Cawangan Kelantan. She is a Chief Executive Officer (CEO) of a charity company, Maab Barakah Resources in Machang, Kelantan. Her company interest in identifying the strategy to solve the poverty issue in Malaysia. She can be contacted at email: niksitimadiah@gmail.com.