Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Themes and Trends in Text Mining Research: Insights from Online News and Annual Reports

Syerina Azlin Md Nasir[1,*], Nik Siti Madihah Nik Mangsor[1], Wan Marhaini Wan Omar[2], Ainul Azila Che Fauzi[3], Au Thien Wan[4]

[1] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Kelantan, Kota Bharu, Malaysia
[2] Faculty of Business and Management, Universiti Teknologi MARA Cawangan Kelantan, Kota Bharu, Malaysia
[3] College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Cawangan Kelantan, Machang, Malaysia
[4] School of Computing and Informatics, Universiti Teknologi Brunei, Brunei

ARTICLE INFO

Keywords:
Annual reports; document clustering; online news; systematic literature review; text mining

ABSTRACT

This study aims to uncover recent trends and themes in text mining research, focusing on online news and annual report data sources. These sources are rich in content, real-world context, and domain-specific information, making them crucial for text mining. The primary research questions guiding this study are: *1) What are the emerging trends in text mining research applied to online news and annual reports? 2) What are the themes generated based on systematic literature review and text mining technique?* To address this question, the study systematically reviews a large number of related studies using the PRISMA review protocol and text mining techniques from the SCOPUS and Web of Science databases. After thorough evaluation, 34 selected articles were analyzed. The PRISMA review protocol ensures transparency and completeness in reporting the review process through its standardized approach for systematic reviews. Additionally, this systematic review explores advancements in text mining techniques such as document clustering and topic modeling, which have facilitated the identification and extraction of relevant evidence from vast amounts of textual data. The study's findings identified four primary themes (text mining/text analytics, machine learning, deep learning, ensemble methods) with 19 sub-themes related to each theme's methodology when applying the PRISMA protocol. By utilizing a text mining technique, five topics were uncovered based on article keywords (text mining, text analytics, machine learning, deep learning, and ensemble) and ten topics emerged based on article abstracts. Underlying both approaches is the consistent recognition of four main areas: text mining/text analytics, machine learning, deep learning, and ensemble methods. This systematic review offers a comprehensive overview of recent text mining research and the emerging trends in this field. It highlights the importance of systematic reviews in synthesizing existing literature and identifying areas for future research.

* *Corresponding author.*
E-mail address: syerina@uitm.edu.my

https://doi.org/10.37934/araset.64.2.165186

1. Introduction

Text mining research has gained significant traction in recent years, driven by the increasing availability of online data as valuable sources. These sources provide a wealth of information that can be leveraged to gain insights and make informed decisions across various domains. The rapid advancement of information technology has led to the emergence of new applications such as social networks and e-commerce, which serve as central hubs for gathering and disseminating information. These platforms generate massive amounts of heterogeneous data, including text from reports, scientific articles, tweets, product reviews, and more [1]. According to [2], Big Data technologies will make unstructured data even more prevalent, with an estimated 150 zettabytes of unstructured data requiring analysis by 2025 for decision-making and predictive analysis. This data is crucial for organizations, aiding in decision-making, predictive analysis, and pattern identification [3]. A notable trend in text mining research is the focus on using online news and annual reports for sentiment analysis, fraud detection, and opinion mining purposes. Unlike previous approaches that primarily relied on quantitative financial information for fraud detection, recent studies have begun to consider qualitative textual content in annual reports to predict fraudulent behaviours [4]. This approach examines the writing and presentation styles in annual reports as valuable indicators of fraud. Furthermore, investigations into detecting deceit in financial statements can contribute to refining general theories of deception. However, handling such unstructured data consistently presents time-consuming and costly challenges for organizations [5]. As the influx of data continues to grow, the associated challenges with its management are also increasing exponentially. The growing complexity of data presents two major challenges: developing approaches for handling massive datasets and addressing the issue of high dimensionality.

To address these challenges, researchers have increasingly turned to text mining techniques to efficiently extract and analyze information from unstructured textual data [6-7]. This trend reflects a growing recognition of the valuable insights that can be gained through systematic analysis of vast amounts of textual information [8,9]. By delving into this treasure trove of data, scientists can uncover patterns, trends, and correlations that may not be immediately apparent through traditional methods alone. The ability to harness the power of text mining in scientific research holds great promise for advancing our understanding across various fields and unlocking new frontiers in innovation and discovery. A critical component driving this shift towards text mining is its potential to enhance the exploration of key thematic questions within diverse subject areas. Whether it's delving into medical literature to glean crucial insights on disease pathways or parsing through environmental reports for emerging sustainability trends, leveraging text mining methodologies allows researchers unprecedented depth and breadth when exploring their chosen themes [10].

Moreover, by effectively navigating through mountains of textual data using advanced computational algorithms fuelled by natural language processing capabilities, scientists can gain an edge in generating evidence-based conclusions that propel further inquiry forward at an accelerated pace [11,12]. Embracing state-of-the-art approaches anchored in rigorous text mining practices empowers scientific communities worldwide with richer contextualized perspectives rooted firmly in empirical findings substantiated upon robust analyses. Therefore, this paper focuses on recent works in the field of text mining research, specifically examining the use of online news and annual reports as data sources. It aims to answer the research questions which are twofold: 1) What are the emerging trends in text mining research applied to online news and annual reports? 2) What are the themes generated based on systematic literature review and text mining technique? This study offers an insightful review of advancements and innovations in this area, highlighting how

researchers leverage these sources to extract meaningful information for decision-making and business intelligence.

2. Methodology

This section discusses the approach used to find articles on text mining techniques utilized by previous researchers. This paper delves into two approaches to confirm the findings: PRISMA review protocol and text mining technique. The reviewers employed PRISMA, which involves data from resources (Scopus and Web of Science) and used to run the systematic review, inclusion and exclusion criteria, steps of the review process (identification, screening, eligibility) and data abstraction and analysis.

2.1 Systematic Literature Review – guided by PRISMA

The review was conducted in accordance with the PRISMA Statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). According to Sierra-Correa and Cantera Kintz, it provides three distinct benefits: 1) formulating clear research questions that enable systematic inquiry, 2) identifying inclusion and exclusion criteria, and 3) attempting to review a large database of scientific literature within a specific timeframe. The PRISMA review protocol facilitates a thorough exploration of terms related to text mining or text analytics. This approach can be utilized for monitoring the adaptation of text mining techniques towards uncovering new patterns in textual datasets.

2.1.1 Formulation of research questions

The formulation of research questions for this study is based on PICo. PICo helps authors to create appropriate research questions for reviews. PICo comprises of three main concepts namely population or problem, interest, and context. In this study, the population can be described as annual reports and online news. Then, it explained the context of analytical techniques such as text mining and text analytics. Based on this concept, research questions are formulated in twofold; “What are the research trends of text mining research in an online news and annual reports”? and “What are the themes generated based on systematic literature review and text mining technique”?

2.1.2 Systematic searching strategies

In systematic searching strategies, there are three processes involved namely identification, screening the inclusion criteria and eligibility (Figure 1).

i. Identification

The assessment used two major bibliographic databases: Scopus and Web of Science. Web of Science is a comprehensive database with over 21,000 journals covering multiple disciplines, including computer science, data mining, big data analytics and other science and technology research domain. It has extensive back file and citation data spanning over 100 years. Clarivate Analytics manages the database and ranks journals based on citations, papers published, and citations per paper. The second database used in the review is Scopus, one of the largest databases for peer-reviewed literature with over 36,000 journals from worldwide publishers. Its subject areas are diverse, including computer science, data mining, big data analytics and others.

The systematic review process consisted of four distinct stages and was conducted in February 2024. Initially, keywords related to text mining, text analytics, social analytics, annual reports, and online news were identified using previous studies and thesaurus as shown in Table 1. Using the keywords, 164 articles were identified from Web of Science and 36 articles from Scopus. Subsequently, careful screening led to the removal of 13 duplicated articles during this phase.

Table 1
The search string used for the systematic review process

Databases	Keywords used
Scopus	TITLE-ABS-KEY ((“Social Analytics" OR "Text Mining" OR "Text Analytics”) AND (“Annual Report" OR "Online News")) AND PUBYEAR > 2018 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (PUBSTAGE , "final"))
Web of Science	ETS= ((“Social Analytics" OR "Text Mining" OR "Text Analytics") AND ("Annual Report" OR "Online News"))

ii. Screening the inclusion criteria

In this stage, several criteria were considered to refine the articles. Firstly, only empirical data articles from academic journals are included, while review articles, books, book chapters, and conference proceedings are excluded. Secondly, non-English publications were omitted to prevent translation issues; the focus is solely on English-language articles. Thirdly, a 5-year period (2019-2023) was selected to allow for observing the development of research and related literature. The review specifically targeted articles in social science-based indexes; thus, excluding articles from hard science indexes such as Science Citation Indexed Expanded. Finally, given its focus on text mining objectives, only articles centred around data sources from online news or annual reports were chosen (refer Table 2). At this stage, out of 187 articles eligible to be reviewed, a total of 147 articles were removed. Hence, 40 articles were selected after the removal process.

Table 2
The inclusion and exclusion criteria

Criterion	Inclusion	Exclusion
Literature type	Journal articles	Proceeding, book, book chapter, review articles.
Language	English	Non-English
Timeline	Between 2019 and 2023	<2019
Indexes	Social Science Citation Index, Emerging Sources Citation Index, Art and Humanities Index (Web of Science)	Science Citation Indexed Expanded (Web of Science)
Data sources	Annual report or Online news news	Not from annual report or online news

iii. Eligibility

Eligibility is the process where the authors manually monitored the selected articles to ensure all the remaining articles (after the screening process) are in line with the criteria. In the eligibility stage, thorough examination led to the exclusion of six articles that did not meet specific criteria. These included being outside the scope of text mining research or not originating from online news, annual reports, or review articles. Finally, after completion of all these stages, a total of 34 articles underwent qualitative analysis (Figure 1).

2.1.3 Quality appraisal

In ensuring the content quality of the remaining articles, two experts were consulted for a quality assessment. Petticrew and Roberts (2006) suggested that experts should rank the remaining articles into three quality categories which are high, moderate, and low. Only articles categorized as high and moderate should be proceeded for review. The experts focused on the methodology of the articles to determine the rank of the quality. For the articles to be included in the review, both authors must mutually agree that the quality must at least be at a moderate level. After thorough discussions among the experts, all the remaining articles were eligible for review.

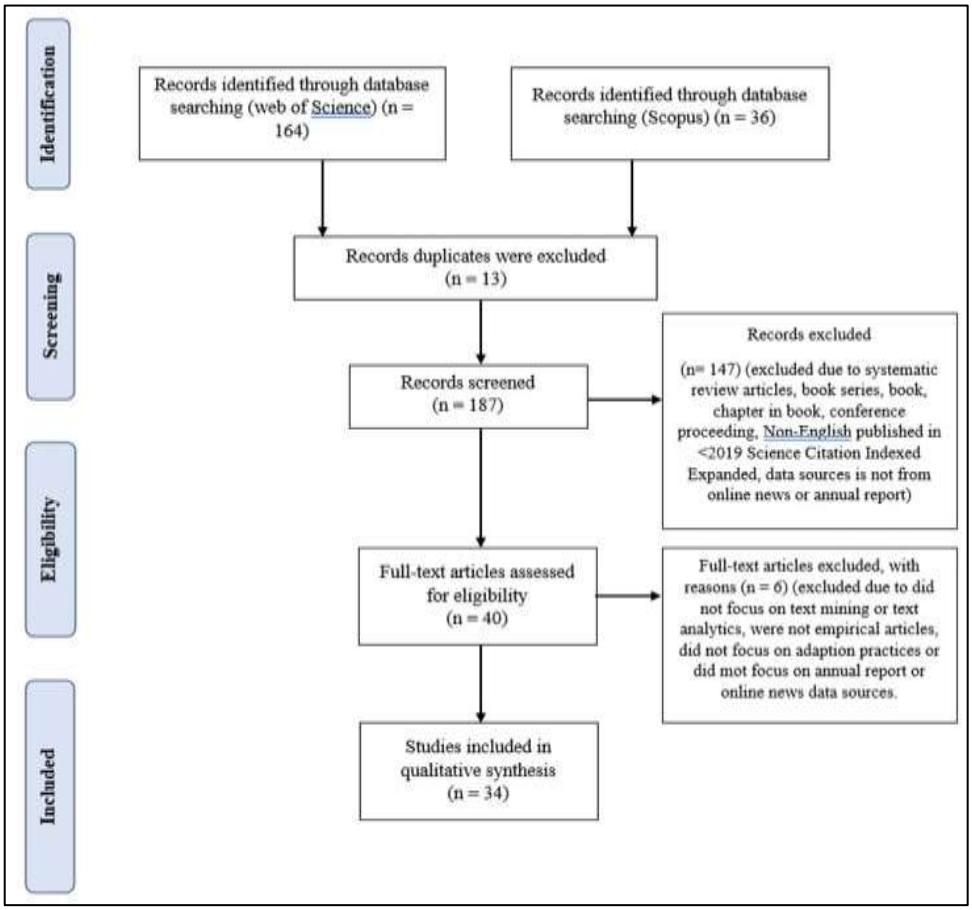

Fig. 1. The flow diagram of the study (Adapted from Shaffril et al., 2018)

2.1.4 Data abstraction and analysis

The remaining articles were evaluated and examined, with emphasis placed on studies that specifically addressed the formulated queries. The data abstraction was conducted based on the research questions, it denotes that any data from the reviewed studies that can answer the research questions were abstracted and placed in a table. Data gathering involved reviewing abstracts first, followed by a thorough reading of the complete articles to identify relevant main topics and subtopics. Qualitative assessment was conducted using content analysis to recognize themes associated with the practices in Text Mining. The authors then structured sub-themes within the typology-based themes.

2.2 Text Mining Approach

This section focuses on the text mining process and is further described below:

2.2.1 Frequency analysis

Published across various domains, abstracts and keywords often consist of unstructured textual data that requires deciphering. Text mining techniques offer an effective means to conduct exploratory analysis and identify semantic patterns. One technique involves visualizing word frequency using a word cloud, which organizes words based on their occurrence in the text. Word clouds are widely used for visually representing textual content and prove valuable for analyzing diverse types of text data [13].

2.2.2 Document clustering and topic modelling

Text clustering is based on the Cluster hypothesis, which states that important texts must share more similarities than non-related ones [14]. Clustering is a reliable approach that is commonly used for analysing massive volumes of data, such as data mining. Text clustering has been demonstrated to be one of the most effective strategies for analysing text themes. Furthermore, it facilitates the topic analysis approach in which named entities with concurrent occurrences are grouped together before being subjected to a clustering process in which frequent items are arranged in sets using the hyper graph-based method [14].

Topic modelling, one of the most widely used text mining techniques, is a methodical and effective way to analyse thousands of documents in a matter of minutes. Latent Dirichlet Allocation (LDA), which is founded upon statistical distributions, is an extensively utilised and valid model within the domain of topic models [15]. LDA operates under the assumption that a correspondence exists between documents and words within a corpus denoted by a bag-of-words. LDA identifies terms that are semantically related and appear in multiple documents of a corpus. These word collections or "topics" are subsequently interpreted as significant "themes" through human intuition [16].

Our documents were abstracts and keywords; henceforth, the terms "keyword" and "document," respectively, were used interchangeably with "abstract" and "document." LDA attributes a probability value to each set of words in relation to each topic, as well as a probability value for each topic in relation to each document. The LDA results for a set of n documents (including abstracts and keywords), m words, and t topics were as follows: the probability ($W_i \mid T_k$) that each word was assigned to a specific topic, and the probability ($T_k \mid D_j$) that each topic was assigned to a document (Figure 2). The terms that ranked highest for each topic in descending P($W_i \mid T_k$) order were utilised

to symbolise the topics. Additionally, we calculated the annual weight of topics using P(T_k | D_j). As an illustration, suppose the initial one hundred abstracts (D_1, D_2,..., D_{100}) were published in 2009. In that case, the weight of T_1 in 2009 would be $\sum^{100}_{j=1} P(T_1|D_j)$ [17].

$$LDA \rightarrow \begin{array}{c} \\ Words \end{array} \begin{array}{c} Topics \\ \begin{bmatrix} P(W_1|T_1) & \dots & P(W_1|T_t) \\ \vdots & \ddots & \vdots \\ P(W_m|T_1) & \dots & P(W_m|T_t) \end{bmatrix} \end{array} \& \begin{array}{c} \\ Topics \end{array} \begin{array}{c} Documents \\ \begin{bmatrix} P(T_1|D_1) & \dots & P(T_1|D_n) \\ \vdots & \ddots & \vdots \\ P(T_t|D_1) & \dots & P(T_t|D_n) \end{bmatrix} \end{array}$$

Fig. 2. Matrix interpretation of LDA.

3. Results

3.1 Trend and Frequency Analysis

Both analyses are used to gain a comprehensive understanding of the textual data using both systematic literature review and text mining approach respectively.

3.1.1 Systematic literature review – trend analysis

Statistical trend analysis of topics can help discover hidden temporal patterns, allowing researchers to go beyond surface-level observations of study trends. Trend analysis for each research area based on year and data sources are shown in the bar graph below (Figure 3 – Figure 6).

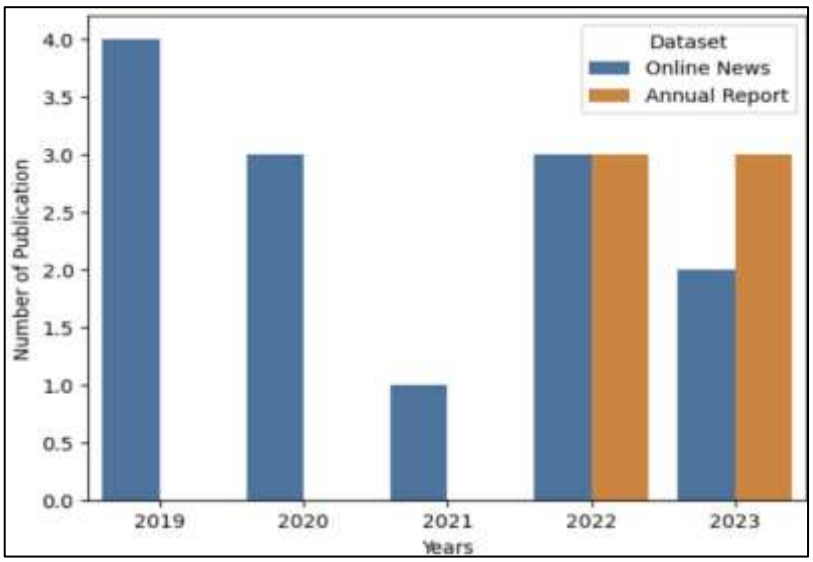

Fig. 3. Number of publications per year based on text mining/text analytics domain area

In text mining/text analytics domain area, Figure 3 reveals a decrease in the number of articles published on online news over the initial three years, followed by an upward trend in the fourth year and subsequent decline in the fifth year. This suggests that the positive trend observed in the fourth year was not sustained. It is also worth noting that no articles were published during the first three years, indicating that researchers only began using annual reports as a data source starting from 2022. The bar graph illustrates a gradual progression from zero publications to a stable number of three publications in both 2022 and 2023. The increased use of annual reports in text mining research since 2022 can be attributed to several factors. Firstly, there is a growing emphasis on corporate sustainability, as regulators and investors demand greater transparency in companies'

environmental, social, and governance disclosures [18] This has led to more comprehensive annual reports, which researchers are leveraging to extract insights into corporate sustainability, risk management, and governance. Secondly, the rise of digital reporting formats has made it easier for researchers to access and analyze large volumes of annual report data and advancements in natural language processing have improved their ability to interpret this unstructured text [19]. Finally, annual reports provide a consistent, longitudinal data source, enabling researchers to study the evolution of corporate strategies, financial health, and narratives over time.

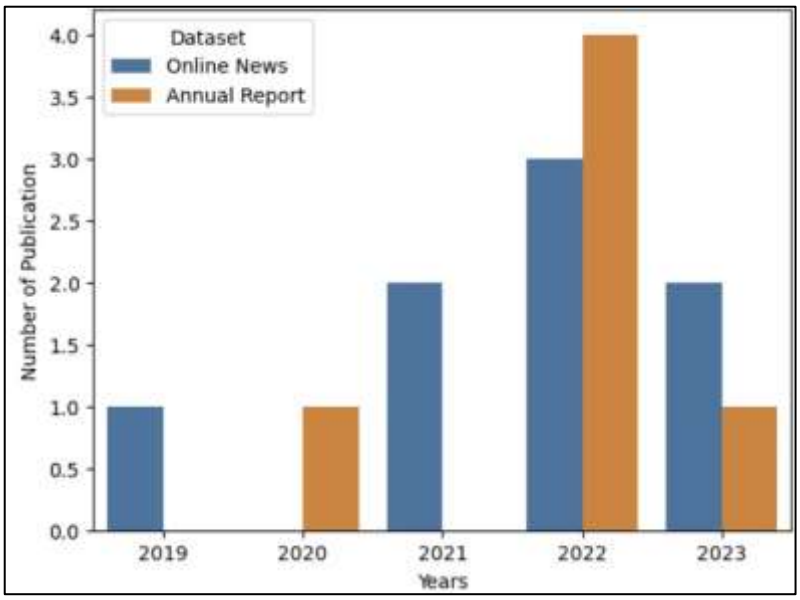

Fig. 4. Number of publications per year based on machine learning domain area

Based on Figure 4, the number of publications on both online news and annual reports in machine learning domain area follows similar patterns, although publication on annual reports only began in 2020. The bar graph showing the number of articles published over five years forms a bell-shaped curve, resembling a normal distribution. In 2022, there was a significant increase leading to the highest peak in article publication. The spike in machine learning publications in 2022 suggests a significant inflection point in the application of machine learning to text mining research. This can be attributed to several key factors: advancements in AI and natural language processing technologies, such as transformer-based models like BERT and GPT, which have revolutionized text analysis; increased accessibility of computational resources and machine learning frameworks, democratizing the use of advanced techniques; and the growing adoption of machine learning across diverse domains, from analyzing news and annual reports to extracting insights from social media and other unstructured data sources. Together, these developments have contributed to the sharp rise in machine learning-related text mining research, reflecting its growing importance as a powerful tool for extracting value from vast amounts of textual data.

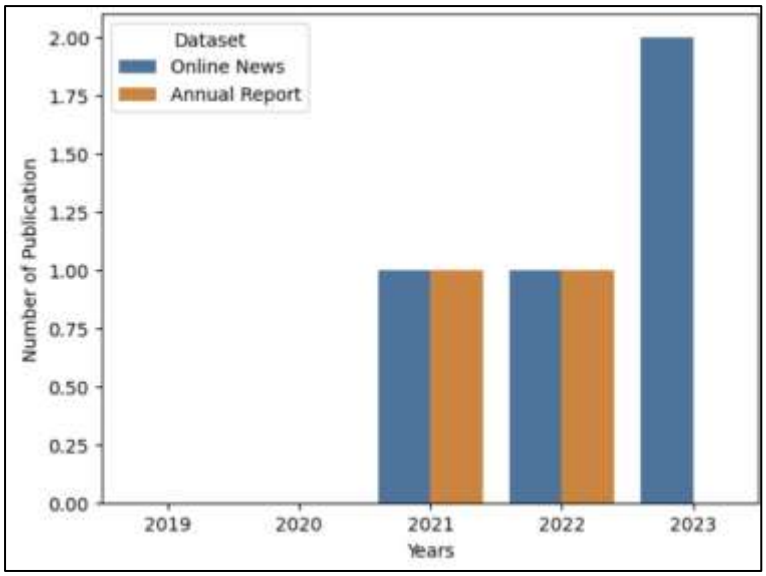

Fig. 5. Number of publications per year based on deep learning domain area

In deep learning domain area as depicted in Figure 5, for both textual data, publication starts in 2021. For online news, there is an increasing trend in article publication. For annual reports, the number of articles published was minimal, only 1 publication each in year 2021 and 2023. The absence of any publications for 2023 is also noteworthy. Overall, the deep learning domain in text mining, as reflected in Figure 5, shows promising growth in the use of online news as a data source, likely due to deep learning's superior ability to handle unstructured, dynamic content. However, the minimal use of annual reports suggests that this structured data may not necessitate the complexity of deep learning models, or that challenges remain in applying deep learning effectively to this domain. The absence of deep learning publications in 2023 indicates a potential shift in research priorities, whether due to the emergence of new models, computational challenges, or environmental concerns related to the heavy use of resources for training deep learning algorithms.

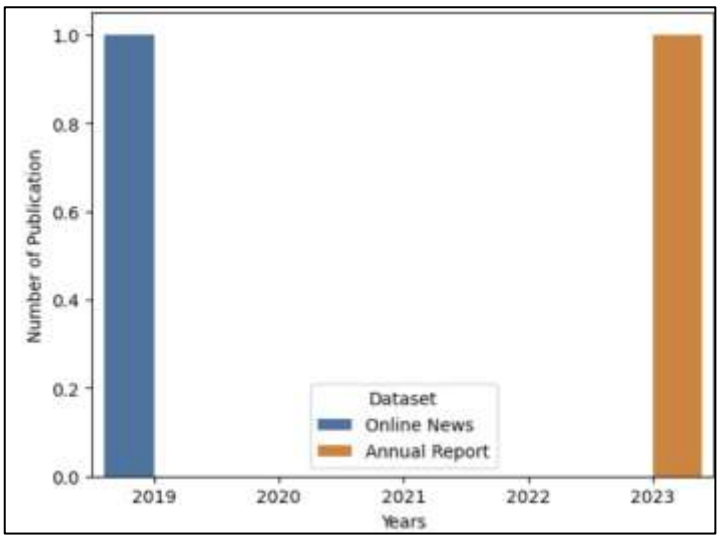

Fig. 6. Number of publications per year based on ensemble method domain area

Figure 6 shows a sparse distribution of articles in ensemble method domain area published over the five-year period. In 2019, only one online news publication was recorded. Similarly, in 2023, there was just one annual report publication. This single instance indicates an attempt to explore ensemble methods in text mining research and suggests potential for further exploration in the future. Ensemble methods, which combine multiple models to improve predictive performance, are often resource-intensive and complex to implement. They require significant computational power and careful tuning of hyperparameters, which can be a barrier for researchers working with limited resources. This may explain the sparse use of ensemble methods in text mining, where simpler models like decision trees or single neural networks may be seen as more feasible for handling large text datasets.

3.1.2 Text mining approach – frequency analysis

To have an overall perspective, the articles were analyzed the frequency of top-10 words based on ‘abstract’ and ‘keywords’ data respectively (refer Table 2 and Table 4) and word cloud. A word cloud represents the frequency of words in a corpus using word size – where larger size denotes higher frequency (Figure 7 and Figure 8).

As per Table 3, we can notice that the most frequent linked words among all the abstract of articles are: “news” followed by “online”, “financial”, “text”, “article”, “study”, “model”, “data”, “information”, and “based” respectively. The prominence of words like "news," "online," and "financial" in the text mining research literature highlights several key trends. First, there is a growing reliance on digital news sources, as online media becomes a crucial source for real-time insights on a range of topics, from financial markets to social issues. Second, the focus on "online" data reflects a shift towards using text mining for real-time or near real-time applications, where staying updated with the latest information is critical. Finally, the high frequency of the word "financial" indicates an increased interest in analyzing financial news and reports, driven by the need to predict market movements, assess company performance, and conduct sentiment analysis on financial narratives. Together, these trends underscore the evolving priorities and applications of text mining research, as researchers leverage diverse digital data sources to extract valuable insights in a timely manner.

Table 3
The Top-10 most frequent words based on ‘abstract’ data

Words	Frequency
news	115
online	54
financial	49
text	45
articles	41
study	39
model	38
data	37
information	33
based	32

Furthermore, as per Table 4, we can notice that the most frequent linked words among all the keyword of articles are: “mining” followed by “text”, “news”, “sentiment”, “learning”, “analysis”, “online”, “data”, “deep”, and “detection” respectively. These results indicate that the most frequent linked words are focused on studies of the implementation of text mining analysis on news data. The

prominent use of keywords like "mining," "sentiment," and "detection" in the text mining research literature suggests a focus on core applications of the field. Text mining is a central research approach, with techniques for extracting patterns, relationships, and topics from large text corpora. Sentiment analysis is a popular application, leveraging text mining to understand public opinion, market sentiment, and customer feedback, particularly in the context of online news. Additionally, text mining is being employed as a predictive tool for event detection and anomaly detection, often utilizing advanced machine learning techniques like deep learning.

Table 4
The Top-10 most frequent words based on 'keyword' data

Words	Frequency
mining	30
text	20
news	14
sentiment	10
learning	9
analysis	9
online	7
data	7
deep	6
detection	6

As shown in Figure 7, we can notice that "news" is the most keyword that was mentioned across all the collected abstract of articles. The second highest frequent words are "online", "financial" and "text" respectively. The dominant presence of "news" and "online" in the word cloud reflects the prevalent focus on online news data as a key source for text mining research. This indicates that the majority of studies are centered around analyzing news articles, likely due to the sheer volume of text and the rich, real-time nature of this data. Online news data is valuable because it provides fresh, unstructured content that can be analyzed for sentiment, topics, and even event detection. However, this raises questions about potential biases introduced by focusing too much on news, such as media slant, the fleeting nature of news articles, and the challenge of generalizing findings across different data types (e.g., social media, forums, or other public datasets).

Fig. 7. Word cloud of high-frequency words based on abstract

Additionally, as shown in Figure 8, we can notice that "mining" is the most keyword that was mentioned across all the collected keyword of articles. The second highest frequent words are "text" and "news" respectively. The increasing number of the words (news, text and mining) could be attributed to the fact that the data sources are from online text news. The prominence of "mining"

alongside “text” reinforces the core theme of text mining as the primary research methodology. This suggests that most studies are deeply focused on methods to extract information, uncover patterns, or classify textual data. As text data continues to proliferate, the importance of mining useful insights from this unstructured data is becoming increasingly critical. The dominance of these terms suggests that text mining remains a central research area, but it also highlights opportunities for the field to evolve by incorporating more complex datasets, such as multimodal (text, images, audio) or heterogeneous data from various sources.

Fig. 8. Word cloud of high-frequency words based on keyword

3.2 Theme Generation

Theme generation enhances the clarity, coherence, and depth of insights derived from a diverse range of scholarly works in text mining as follows:

3.2.1 Systematic literature review – theme generation

The researcher performed thematic analysis to identify themes and sub-themes. Thematic analysis is considered as the most suitable in synthesizing a mixed research design. The result yielded five primary themes, and 19 sub-themes related to each theme's methodology. The five main themes are text mining/text analytics (with eight sub-themes), machine learning (with five sub-themes), deep learning (with three sub-themes) and ensemble methods (with three sub-themes) as shown in Table 5. These results present a comprehensive examination of the current methodological approaches utilized by previous researchers. A total of twenty-two studies focused on online news sources [20-39), while twelve studies concentrated on companies' annual reports (40-51). In terms of publication years, nine articles were published in 2023, twelve in 2022, three in 2021, four in 2020, and six in 2019.

Table 5
The findings

AUTHOR	DATA SOURCE		TEXT MINING/ TEXT ANALYTICS								MACHINE LEARNING					DEEP LEARNING			ENSEMBLE		
	ON	AR	IE	IR	TTD	CC	CL	NLP	IV	SM	R	C	CL	DAN	VDR	D	G	HLO	BAGGING	BOOSTING	STACKING
Xiaoyu & Shengyong (2022).		/															/				
Suh. (2019).	/																			/	
Li et al. (2021).		/															/				
Oh et al. (2023).	/							/													
Guo & Xing. (2023).		/						/													
Chi et al. (2022).		/										/									
Huang & Wei (2022).		/						/			/										
Mao et al. (2022).		/						/			/										
Olson & Chae (2023).		/			/																
Arianto et al. (2021).	/						/					/		/							
Kim & Cho (2022).		/						/													
Yang et al. (2022).		/											/								
Bharathi & Geetha (2019).	/							/													
Majumdar & Singh (2023).		/						/													
Cui et al. (2023).	/				/							/									
Mohsin et al. (2019).	/							/													
Mohaimin et al. (2023).	/															/					
Bechini et al. (2022).	/											/									
Lee & Park (2019).	/							/													
Christensen et al. (2022).	/							/													
Jin & Joo (2020).	/							/													
Cui et al. (2023).	/											/									
Qian et al. (2019).	/												/								
Wu et al. (2023).	/															/					
Hossain et al. (2021).	/											/				/					
Chu et al. (2020).	/							/													
Chen & Chen (2019).	/							/													
Yan et al. (2022).	/															/					
Tang & Wei (2023).		/										/								/	
Jung et al. (2022).	/			/								/									
Zhang et al. (2020).	/			/																	
Shaikh et al. (2022).	/							/													
Craia et al. (2020).		/									/										
Bechini et al. (2022).	/											/									

IE	Information Extraction	NLP	Natural Language Processing	DAN	Detection of Anomalies and Novelties
IR	Information retrieval	IV	Information Visualization	VDR	Visualization and Dimensionality Reduction
TTD	Topic Tracking and Detection	SM	Summarization	D	Discriminative
CC	Categorization/ Classification	R	Regression	G	Generative
CL	Clustering	C	Classification	HLO	Hybrid Learning and Others
ON	Online Newspaper	AR	Annual Report		

3.2.2 Text mining approach – theme generation

Beyond preliminary frequency assessments, we used LDA to disclose the hidden semantic structure of papers. LDA has an important pre-processing step, which is defining the optimal number of topics. We used the latent concept modelling to estimate the optimum point. This model maximizes the overall dissimilarity between the word distributions of topics.

Using the Python Gensim package, we found the optimum number of topics at 10 by applying the latent concept modelling on the number of topics from 2 to 10. After removing stop words (e.g., ‘‘the’’ and ‘‘a’’), we applied the Variational Bayes Inference algorithm implementation of LDA with its default settings on the abstracts. Table 6 outlines the 10 topics identified through a Latent Dirichlet Allocation analysis of abstract data from text mining studies. The topics were systematically named based on their top 10 most frequent and relevant terms, enabling a clear interpretation of the dominant themes present in the abstract corpus.

Table 6
The 10 topics of text mining studies based on 'abstract' data generated by LDA

Topic	Name	Top 10 Words
Topic 1	Sentiment Analysis	'news', 'article', 'enterprise', 'effect', 'sentiment', 'intelligent', 'development', 'medium', 'model', 'language'
Topic 2	Information Extraction	'financial', 'emergency', 'model', 'event', 'chinese', 'fraud', 'extraction', 'textual', 'using', 'statement'
Topic 3	Clustering	'sentiment', 'study', 'news', 'financial', 'city', 'lexicon', 'model', 'cluster', 'change', 'based'
Topic 4	Information Retrieval	'volatility', 'word', 'text', 'knowledge', 'embedding', 'financial', 'method', 'model', 'prediction', 'learning'
Topic 5	Natural Language Processing	'model', 'financial', 'result', 'annual_report', 'feature', 'ratio', 'fraud', 'sentence', 'process', 'statement'
Topic 6	Text Classification	'news', 'vaccine', 'online', 'article', 'event', 'covid', 'business', 'result', 'term', 'information'
Topic 7	Topic Tracking & Detection	'news', 'model', 'article', 'financial', 'risk', 'online', 'disease', 'study', 'problem', 'social'
Topic 8	Information Visualization	'risk', 'firm', 'document', 'pandemic', 'data', 'covid', 'analysis', 'technology', 'news', 'using'
Topic 9	Summarization	'time', 'news', 'online', 'keywords', 'newspaper', 'agro', 'found', 'term', 'different', 'daily'
Topic 10	ML Classification	'esports', 'text', 'data', 'news', 'online', 'asian_game', 'sport', 'image', 'gender_equality', 'paper'

The detected 10 topics was named, for example, Topic 1 is was named **Sentiment Analysis** due to the prominence of terms such as "sentiment," "news," and "effect," which suggest a focus on the examination of sentiment within textual data, particularly in the context of news articles and the impacts of sentiment on organizations. Sentiment analysis is a widely utilized technique for evaluating emotions or viewpoints present in textual content. When applied to news articles and social media, it can offer insights into public opinion, corporate strategies, and the effects of media. The emphasis on words like "intelligent," "development," and "enterprise" implies an interest in the commercial and practical applications of sentiment analysis, such as monitoring consumer feedback or corporate sentiment.

Topic 2 focuses on the field of **Information Extraction**, as evident from the prominence of terms like "extraction," "financial," and "textual." The research in this area centres around the process of retrieving structured data from unstructured textual sources, particularly in the domains of financial reporting, emergency event analysis, and fraud detection. The inclusion of the word "Chinese" suggests that some of the studies investigate datasets from China, with a likely emphasis on financial information. Information extraction is a crucial technique for various applications, such as search engines, financial forecasting, and legal analysis. The research in this topic leverages predictive text mining methods, often in combination with advanced text embedding techniques, to enhance the accuracy and efficiency of information retrieval from large-scale data repositories.

Document Clustering is one of the most important text mining technologies, intended to assist users in successfully navigating, summarizing, and organizing text materials. The prominent use of terminology such as "cluster," "sentiment," and "model" implies that this topic centres on the application of clustering methods to analyze financial news and other textual datasets. The emphasis on "financial," "sentiment," and "news"-related clustering suggests that a significant portion of the research in this domain focuses on organizing large-scale financial or news-oriented datasets to enhance interpretability. Clustering techniques serve to reduce the complexity inherent in analyzing

extensive text corpora by grouping together similar documents, thereby facilitating more effective extraction of insights from unstructured data, particularly in fields such as finance and media.

Information Retrieval appears to focus on the retrieval of relevant information from large text datasets, potentially for applications in financial forecasting or predictive modeling. The keywords "volatility" and "embedding" suggest the utilization of techniques such as text embedding and machine learning models to enhance the quality and accuracy of information retrieval. Information retrieval is a fundamental process of extracting relevant data from extensive datasets. In this context, the theme seems to encompass predictive modeling and knowledge extraction, with a specific emphasis on financial datasets. The inclusion of the term "embedding" indicates the employment of advanced text representation methods, like word embeddings, to improve the effectiveness of information retrieval and search outcomes.

Topic 5 was labeled as **Natural Language Processing (NLP)** due to the presence of key terminology like "model," "sentence," and "process," which are fundamental concepts in the field. The inclusion of words such as "financial," "fraud," and "annual_report" suggests the application of NLP techniques to analyze formal documents, particularly financial reports. NLP is a crucial tool in industries like finance, where the capacity to process and comprehend large volumes of reports, statements, and regulatory filings can significantly impact decision-making and fraud detection.

Text Classification is the process of categorising materials according to their content. This topic focuses on the classification of textual data, particularly pertaining to online news articles and COVID-19-related information. The presence of terms like "news," "vaccine," and "covid" suggests a concentration on categorizing and analyzing such text-based content. Furthermore, the inclusion of "business" and "event" indicates applications in real-time monitoring of events and business-oriented analyses. Text classification is a fundamental technique in the field of text mining, with widespread use cases in news topic categorization, medical research, and tracking sentiments for business purposes.

Topic Monitoring and Detection is focused on the identification and monitoring of emerging themes and trends in real-time, particularly in the contexts of financial risks, public health crises, and social developments. The prominence of terms like "topic," "tracking," and "detection" suggests a research emphasis on techniques for rapidly identifying and following key issues as they unfold, which is of crucial importance for domains that require timely responses, such as healthcare and financial risk mitigation.

Topic 8 is labelled as **Information Visualization** which focuses on the visualization of information, particularly in the contexts of risk management and the COVID-19 pandemic. The prominence of terms like "risk," "pandemic," "data," and "visualization" suggests a research emphasis on presenting data-driven insights through graphical methods. Data visualization plays a crucial role in aiding decision-makers by transforming complex datasets into more comprehensible forms, which is especially important in domains such as risk management, public health crises, and financial reporting.

Text Summarization is the process of reducing a document's length and complexity while maintaining the most important elements and general meaning. The top words in this topic like "time," "keywords," and "online" point to studies that involve summarizing large datasets, particularly news articles. This topic likely addresses the creation of concise summaries from large volumes of text, such as newspapers or daily news sources. Summarization helps extract key information from vast amounts of text quickly, which is crucial for news platforms, media analysis, and organizations that need to monitor daily developments in their fields.

The last topic is named after **Machine Learning (ML) Classification**. This topic includes words like "esports," "text," and "gender_equality," suggests a focus on applying machine learning techniques

to classify and analyze diverse textual data, ranging from topics related to esports, gender equality, and other societal issues. The research in this area appears to explore the intersection of machine learning and text classification, leveraging these methods to gain insights from large datasets across various domains, including sports, media, and social contexts.

The topic labels were generated by identifying the most frequently occurring and conceptually salient terms within each cluster. This approach ensured that the labels accurately captured the primary focus of the articles in the respective clusters. The prevalence of terms such as "news," "model," and "financial" across multiple topics indicates that these themes are particularly prominent in text mining research pertaining to online news content and corporate annual reports, underscoring the broad significance of these domains in the analyzed studies. Each topic label provides insight into both the specific applications and the methodological approaches employed in the examined research.

Subsequently, we applied the same methodology to the keywords column. Using the Python Gensim library, we determined the optimal number of topics to be 5 by conducting latent concept modeling on topic counts ranging from 2 to 10. After removing stop words (e.g., ‘‘the’’ and ‘‘a’’), we implemented the Variational Bayes Inference algorithm for LDA with its default parameters to identify the 5 key topics within the keyword data as shown in Table 7.

Table 7
The 5 to7pics of text mining studies based on ‘keywords’ data generated by LDA

Topic	Name	Top 10 Words
Topic 1	Text Mining	'mining', 'text', 'news', 'data', 'medium', 'city', 'covid', 'online', 'sentiment', 'learning'
Topic 2	Text Analytics	'mining', 'text', 'analysis', 'sentiment', 'learning', 'news', 'deep', 'detection', 'analytics', 'method'
Topic 3	Machine Learning	'text', 'mining', 'news', 'online', 'concept', 'embedding', 'word', 'extraction', 'event', 'detection'
Topic 4	Deep Learning	'information', 'conditional', 'random', 'directional', 'term', 'batch', 'blstm', 'trained', 'long', 'word'
Topic 5	Ensemble Method	'clustering', 'covid', 'part', 'random', 'short', 'speech', 'tagging', 'trained', 'area', 'city'

The detected 5 topics was named, for example, Topic 1 is **Text Mining** based on interpretation of ‘‘mining, text, news, data, medium, city, covid, online, sentiment, learning”. These word lists or ‘‘topics’’ are then interpreted by human intuition as meaningful ‘‘themes’’. In this study, text mining and text analytics is considered the same. Therefore, there are four main themes derived. This further confirmed that the research areas mostly conducted by researchers are based on the techniques of text mining/text analytics, machine learning, deep learning and ensemble method, respectively. The explanation of each theme is as follow:

i. Text Mining/Text Analytics

A total of 19 out of 34 studies focused on text mining/ text analytics techniques to efficiently extract and analyze information from unstructured textual data. The most common text mining applied are natural language processing (NLP) (14 studies) while 2 studies on information retrieval, another 2 studies applied topic tracking and detection. This unified topic represents the core methodologies utilized to extract meaningful patterns, trends, and insights from text-based data sources. The prominent keywords emphasize the application of these techniques across diverse

domains, with a particular focus on the analysis of news articles and online content. The inclusion of terms like 'covid,' 'sentiment,' and 'learning' suggests that sentiment analysis and machine learning models are extensively employed to examine significant global occurrences, such as the COVID-19 pandemic. The application of text mining and text analytics for the real-time analysis of news and social media data is crucial for understanding public sentiment and behavioral dynamics during such impactful events.

ii. Machine Learning

A total of 14 out of 34 studies implemented machine learning techniques on text dataset. The most common machine learning technique applied are classification (9 studies) while the others applied clustering, regression and detection of anomalies and novelties, respectively. This topic is centred on Machine Learning models and their applications, including their use in prediction tasks and algorithm development. The prominence of terms like 'embedding,' 'word,' and 'extraction' indicates that this topic centers on machine learning methods, particularly those involving word embeddings and information extraction. The occurrence of 'event' and 'detection' suggests the exploration of event detection applications, such as identifying patterns and events from online news and textual data sources. The presence of 'deep' implies that deep learning is a sub-area of interest within this topic. Furthermore, the frequent mention of 'mining' and 'news' suggests the integration of machine learning approaches with text mining to enhance analytical capabilities and predictive modeling.

iii. Deep Learning

A total of 6 out of 34 studies implemented deep learning techniques to find the desire text information. The most common deep learning technique applied are discriminative implementation (6 studies) while two studies on generative technique. The prominent keywords in Topic 4 associated with this topic directly correlate with deep learning methodologies. Terms such as 'blstm,' 'trained,' and 'batch' indicate a focus on sequence learning and model training techniques. The significance of deep learning lies in its enhanced capability to process and comprehend vast volumes of unstructured textual data, outperforming traditional machine learning approaches. Specifically, models like Bidirectional Long Short-Term Memory are widely utilized for tasks such as sentiment analysis, text generation, and sequence prediction, where understanding context and sequential information is crucial. This is a clear demonstration of the application of deep learning models in natural language processing tasks, including classification, language modeling, and sequence prediction. Deep learning is a more advanced subfield of machine learning that brings it closer to artificial intelligence. It allows for the modelling of complicated relationships and concepts at several levels of representation.

iv. Ensemble Method

Two out of 34 studies utilized the boosting ensemble method, resulting in significantly improved performance. This topic indicates a focus on ensemble methods, which involve the combination of multiple learning models to enhance prediction accuracy and robustness. The prominence of keywords such as 'clustering,' 'random,' and 'trained' point to the utilization of ensemble techniques, which often incorporate methods like Random Forests. The presence of 'Covid' and 'speech' suggests applications in public health and speech processing tasks. Ensemble approaches are particularly useful for integrating diverse models to handle complex text mining challenges, as evidenced by the

focus on tasks such as event detection and concept extraction. The increasing adoption of ensemble methods in text mining research can be attributed to their superior performance in tackling complex analytical tasks, including sentiment detection and fraud analysis.

4. Discussion

The research findings reveal key trends with broad implications for both future research and practical applications of text mining. The growing prominence of text mining and text analytics indicates an increasing need for efficient tools to extract, process, and analyze vast amounts of unstructured data from various sources, such as financial reports, social media, and news articles. As organizations continue to generate massive amounts of data, future research will likely focus on developing more advanced algorithms to handle this complexity. This trend suggests a shift towards more automated and scalable solutions, where real-time analysis becomes the norm. Practically, businesses and policymakers can leverage these techniques to gain deeper insights into market trends, consumer behavior, and public sentiment, improving decision-making processes across industries.

The emphasis on machine learning and deep learning underscores the growing importance of adaptive models that can manage and learn from diverse datasets. These methods allow researchers to build predictive models that improve over time, addressing the challenges posed by noisy and unstructured data. Future research in this area may explore hybrid models combining both traditional statistical techniques and machine learning algorithms to enhance performance, particularly in sectors like finance, healthcare, and cybersecurity. In practice, businesses can use machine learning for everything from sentiment analysis to detecting fraudulent transactions, offering real-time solutions that adapt to evolving conditions.

The integration of ensemble methods—combining multiple algorithms to improve accuracy—points to a critical direction for future research. The success of ensemble approaches suggests that single-algorithm models may no longer be sufficient for handling the complexity and variability of large datasets. Future studies may focus on refining these ensemble techniques, improving interpretability, and reducing computational costs. Practically, this development has significant implications for industries where precision is critical, such as risk management, where ensemble methods can provide more accurate predictions of market volatility, disease outbreaks, or customer churn.

One of the most transformative trends is the shift towards multimodal data analysis, which combines structured and unstructured data sources. This trend offers a broader, more holistic understanding of topics like corporate behavior or market sentiment. Future research will likely explore methods for integrating diverse data types—from textual reports to real-time data feeds—addressing the technical challenges of aligning and processing these sources. Practically, this opens new avenues for comprehensive market analyses, public health monitoring, and real-time crisis response, where multiple data streams must be processed simultaneously to inform decisions.

Information visualization is another emerging focus, highlighting the need for more intuitive and interactive tools to represent complex data insights. As datasets grow in size and complexity, traditional quantitative analysis methods may struggle to communicate findings effectively. Future research could concentrate on creating advanced visual interfaces that allow non-technical users to explore and interact with large datasets, enhancing decision-making processes. For practitioners in domains like policy development or business intelligence, effective data visualization will be crucial in translating complex analytical outputs into actionable insights, democratizing access to data-driven decision-making.

The increasing application of text mining in real-time use cases, such as public health monitoring and misinformation detection, suggests a growing role for these technologies in addressing urgent societal challenges. Future research may focus on refining algorithms to handle real-time data streams with greater accuracy and speed. The practical implications are profound, as governments, businesses, and organizations can use these technologies to respond more effectively to crises, track public sentiment during critical events, and combat the spread of misinformation in digital media.

Overall, these trends indicate that text mining will continue to evolve into a critical tool for both academia and industry, with future research likely to focus on improving algorithmic accuracy, scalability, and real-time application. The integration of multiple data sources and the development of more user-friendly visualization tools will ensure that the insights gained through text mining become more accessible and actionable across various sectors.

5. Conclusions

This study has systematically examined and analyzed the trends and prominent themes in text mining research, particularly in the realms of online news and corporate annual reports. Through a comprehensive evaluation of 34 studies using the PRISMA review protocol, the research uncovers the most prominent techniques in text mining, including natural language processing, classification, clustering, and ensemble methods. These findings demonstrate the critical role these methodologies play in extracting valuable insights from large volumes of unstructured textual data, especially in fields like finance, public health, and social media. The systematic approach to identifying these themes has led to a clear understanding of how text mining has evolved and its growing application in real-world scenarios.

The study reveals that the application of machine learning, deep learning, and ensemble techniques are dominant in contemporary text mining research. The combination of models using ensemble methods has proven to enhance accuracy and effectiveness, particularly in dealing with unstructured data, while clustering and topic modeling remain essential for organizing vast datasets. The study also highlights that NLP continues to be a foundational approach, providing significant advancements in areas like sentiment analysis, topic detection, and information retrieval.

The broader implications of these findings suggest that text mining is becoming increasingly crucial for real-time data analysis and decision-making across various industries. As text mining techniques evolve, future research should focus on refining and optimizing existing methods to enhance their scalability and efficiency, especially for large-scale datasets. Additionally, there is potential for integrating multimodal data—combining text with images, videos, or audio—which could offer even richer insights and deeper analysis.

In practical terms, the study's findings point to the growing reliance on text mining techniques in sectors that require sophisticated analysis, such as financial forecasting, crisis management, sentiment analysis, and public policy decision-making. This review has not only identified the current trends but also provided a roadmap for future advancements in the field, encouraging further research to explore new algorithms and applications that could push the boundaries of text analytics. Overall, this study underscores the vital role text mining will continue to play in extracting actionable insights from dynamic and complex datasets, ultimately supporting the growing demand for data-driven decision-making in a rapidly evolving digital landscape.

Acknowledgement
This research was funded by a grant from the Ministry of Higher Education of Malaysia (FRGS Grant: FRGS/1/2023/ICT06/UITM/02/1). We are also grateful for the support provided by Universiti Teknologi MARA.

References
[1] Gonçalves, Carlos Adriano, Adrián Seara Vieira, Célia Talma Gonçalves, Rui Camacho, Eva Lorenzo Iglesias, and Lourdes Borrajo Diz. "A novel multi-view ensemble learning architecture to improve the structured text classification." Information 13, no. 6 (2022): 283. https://doi.org/10.3390/info13060283

[2] Baviskar, Dipali, Swati Ahirrao, Vidyasagar Potdar, and Ketan Kotecha. "Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions." IEEE Access 9 (2021): 72894-72936. https://doi.org/10.1109/ACCESS.2021.3072900

[3] Adnan, Kiran, Rehan Akbar, and Khor Siak Wang. "Information extraction from multifaceted unstructured big data." International Journal of Recent Technology and Engineering (IJRTE) 8 (2019): 1398-1404. https://doi.org/10.35940/ijrte.B1074.0882S819

[4] Humpherys, Sean L., Kevin C. Moffitt, Mary B. Burns, Judee K. Burgoon, and William F. Felix. "Identification of fraudulent financial statements using linguistic credibility analysis." Decision Support Systems 50, no. 3 (2011): 585594. https://doi.org/10.1016/j.dss.2010.08.009

[5] Barnett, Michael L., Andrew Wilcock, J. Michael McWilliams, Arnold M. Epstein, Karen E. Joynt Maddox, E. John Orav, David C. Grabowski, and Ateev Mehrotra. "Two-year evaluation of mandatory bundled payments for joint replacement." New England Journal of Medicine 380, no. 3 (2019): 252-262. https://doi.org/10.1056/NEJMsa1809010

[6] Sundaram, Girish, and Daniel Berleant. "Automating systematic literature reviews with natural language processing and text mining: A systematic literature review." In International Congress on Information and Communication Technology (2023): 73-92. https://doi.org/10.1007/978-981-99-3243-6_7

[7] Hassani, Hossein, Christina Beneki, Stephan Unger, Maedeh Taj Mazinani, and Mohammad Reza Yeganegi. "Text mining in big data analytics." Big Data and Cognitive Computing 4, no. 1 (2020): 1. https://doi.org/10.3390/bdcc4010001

[8] Morshidi, Azizan, Noor Syakirah Zakaria, Mohammad Ikhram Mohammad Ridzuan, Rizal Zamani Idris, Azueryn Annatassia Dania Aqeela, and Mohamad Shaukhi Mohd Radzi. "Artificial Intelligence and Islam: A BibiliometricThematic Analysis and Future Research Direction." Semarak International Journal of Machine Learning 1, no. 1 (2024): 41-58. https://doi.org/10.37934/sijml.1.1.4158

[9] Hashim, Mohd Ekram Alhafis, Nur Safinas Albakry, Wan Azani Mustafa, Banung Grahita, Miharaini Md Ghani, Hafizul Fahri Hanafi, Suraya Md Nasir, and Catherina ana Ugap. "Understanding the Impact of Animation Technology in Virtual Reality: A Systematic Literature Review." International Journal of Computational Thinking and Data Science 1, no. 1 (2024): 53-65. https://doi.org/10.37934/CTDS.1.1.5365

[10] Thakur, Khusbu, and Vinit Kumar. "Application of text mining techniques on scholarly research articles: Methods and tools." New Review of Academic Librarianship 28, no. 3 (2022): 279-302. https://doi.org/10.1080/13614533.2021.1918190

[11] Dogra, Varun, Sahil Verma, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi, and Muhammad Fazal Ijaz. "A complete process of text classification system using state-of-the-art NLP models." Computational Intelligence and Neuroscience 2022 (2022). https://doi.org/10.1155/2022/1883698

[12] Albalawi, Rania, Tet Hin Yeap, and Morad Benyoucef. "Using topic modeling methods for short-text data: A comparative analysis." Frontiers in artificial intelligence 3 (2020): 42. https://doi.org/10.3389/frai.2020.00042

[13] Jayashankar, Shailaja, and R. Sridaran. "Superlative model using word cloud for short answers evaluation in eLearning." Education and Information Technologies 22, no. 5 (2017): 2383-2402. https://doi.org/10.1007/s10639-016-9547-0

[14] Salloum, Said A., Mostafa Al-Emran, Azza Abdel Monem, and Khaled Shaalan. "Using text mining techniques for extracting information from research articles." Intelligent natural language processing: Trends and Applications (2018): 373-397. https://doi.org/10.1007/978-3-319-67056-0_18

[15] Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey." Multimedia tools and applications 78 (2019): 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

[16] Kherwa, Pooja, and Poonam Bansal. "Topic modeling: a comprehensive review." EAI Endorsed transactions on scalable information systems 7, no. 24 (2019).

[17] Karami, Amir, Morgan Lundy, Frank Webb, and Yogesh K. Dwivedi. "Twitter and research: A systematic literature review through text mining." IEEE access 8 (2020): 67698-67717. https://doi.org/10.1109/ACCESS.2020.2983656
[18] Momchilov, G. (2022, March 28). Sustainability reporting by companies: reasons and financial benefits. , 7(1), 55-68. https://doi.org/10.38188/2534-9228.22.1.06
[19] Luccioni, A., Baylor, E., & Duchêne, N. (2020, January 1). Analyzing Sustainability Reports Using Natural Language Processing. Cornell University.
[20] Suh, Jong Hwan. "SocialTERM-Extractor: Identifying and predicting social-problem-specific key noun terms from a large number of online news articles using text mining and machine learning techniques." Sustainability 11, no. 1 (2019): 196. https://doi.org/10.3390/su11010196
[21] Oh, Hyelim, Khim-Yong Goh, and Tuan Q. Phan. "Are you what you tweet? The impact of sentiment on digital news consumption and social media sharing." Information Systems Research 34, no. 1 (2023): 111-136. https://doi.org/10.1287/isre.2022.1112
[22] Arianto, Rakhmat, SW Harco Leslie, Yaya Heryadi, and E. Abdurachman. "Fake news detection model based on credibility measurement for Indonesian online news." J. Theor. Appl. Inf. Technol 15, no. 7 (2021).
[23] Bharathi, SV Shri, and Angelina Geetha. "Determination of news biasedness using content sentiment analysis algorithm." Indonesian Journal of Electrical Engineering and Computer Science 16, no. 2 (2019): 882-889. https://doi.org/10.11591/ijeecs.v16.i2.pp882-889
[24] Cui, Linjie, Eun Joung Kim, and JungYoon Kim. "How Chinese Media Addresses Esports Issues: A Text Mining Comparative Analysis of Online News and Viewers' Comments on the Hangzhou
[25] Mohsin, Mohamad Farhan Mohamad, Siti Sakira Kamaruddin, Fadzilah Siraj, Hamirul Aini Hambali, and Mohammed Ahmed Taiye. "Investigating the relevant agro food keyword in Malaysian online newspapers." Int. J. Adv. Sci. Eng. Inf. Technol 9, no. 6 (2019): 2166-2175. https://doi.org/10.18517/ijaseit.9.6.7955
[26] Mohaimin, Izzati, Rosyzie A. Apong, and Ashrol R. Damit. "Part-of-Speech (POS) Tagging for Standard Brunei Malay: A Probabilistic and Neural-Based Approach." Journal of Advances in Information Technology 14, no. 4 (2023). https://doi.org/10.12720/jait.14.4.830-837
[27] Bechini, Alessio, Alessandro Bondielli, José Luis Corcuera Bárcena, Pietro Ducange, Francesco Marcelloni, and Alessandro Renda. "A news-based framework for uncovering and tracking city area profiles: assessment in Covid19 setting." ACM Transactions on Knowledge Discovery from Data (TKDD) 16, no. 6 (2022): 1-29. https://doi.org/10.1145/3532186
[28] Lee, Young-Joo, and Ji-Young Park. "Emerging gender issues in Korean online media: A temporal semantic network analysis approach." Journal of Contemporary Eastern Asia 18, no. 2 (2019): 118-141.
[29] Christensen, Bente, Daniel Laydon, Tadeusz Chelkowski, Dariusz Jemielniak, Michaela Vollmer, Samir Bhatt, and Konrad Krawczyk. "Quantifying changes in vaccine coverage in mainstream media as a result of the COVID-19 outbreak: Text mining study." JMIR infodemiology 2, no. 2 (2022): e35121. https://doi.org/10.2196/35121
[30] Jin, Hoon, and Dong-Won Joo. "Method and Steps for Diagnosing the Possibility of Corporate Bankruptcy Using Massive News Articles." IEIE Transactions on Smart Processing & Computing 9, no. 1 (2020): 13-21. https://doi.org/10.5573/IEIESPC.2020.9.1.013
[31] Qian, Yu, Xiongwen Deng, Qiongwei Ye, Baojun Ma, and Hua Yuan. "On detecting business event from the headlines and leads of massive online news articles." Information Processing & Management 56, no. 6 (2019): 102086. https://doi.org/10.1016/j.ipm.2019.102086
[32] Wu, Binrong, Lin Wang, Sheng-Xiang Lv, and Yu-Rong Zeng. "Forecasting oil consumption with attention-based IndRNN optimized by adaptive differential evolution." Applied Intelligence 53, no. 5 (2023): 5473-5496. https://doi.org/10.1007/s10489-022-03720-z
[33] Hossain, Arafat, Md Karimuzzaman, Md Moyazzem Hossain, and Azizur Rahman. "Text mining and sentiment analysis of newspaper headlines." Information 12, no. 10 (2021): 414. https://doi.org/10.3390/info12100414
[34] Chu, Chih-Yuan, Kijung Park, and Gül E. Kremer. "A global supply chain risk management framework: An application of text-mining to identify region-specific supply chain risks." Advanced Engineering Informatics 45 (2020): 101053 https://doi.org/10.1016/j.aei.2020.101053.
[35] Chen, Mu-Yen, and Ting-Hsuan Chen. "Modeling public mood and emotion: Blog and news sentiment and socioeconomic phenomena." Future Generation Computer Systems 96 (2019): 692-699. https://doi.org/10.1016/j.future.2017.10.028
[36] Yan, Jianzhuo, Lihong Chen, Yongchuan Yu, Hongxia Xu, Qingcai Gao, Kunpeng Cao, and Jianhui Chen. "Emergeventmine: End-to-end chinese emergency event extraction using a deep adversarial network." ISPRS International Journal of Geo-Information 11, no. 6 (2022): 345. https://doi.org/10.3390/ijgi11060345
[37] Jung, Dongin, Misuk Kim, and Yoon-Sik Cho. "Detecting documents with inconsistent context." IEEE Access 10 (2022): 98970-98980. https://doi.org/10.1109/ACCESS.2022.3204151

[38] Zhang, Yiding, Motomu Ibaraki, and Franklin W. Schwartz. "Disease surveillance using online news: Dengue and zika in tropical countries." Journal of Biomedical Informatics 102 (2020): 103374. https://doi.org/10.1016/j.jbi.2020.103374
[39] Shaikh, Anoud, Naeem Ahmed Mahoto, and Mukhtiar Ali Unar. "TextGraph-A lexicon based framework for concept extraction and visualization." Journal of Intelligent & Fuzzy Systems 43, no. 2 (2022): 2035-2044. https://doi.org/10.3233/JIFS-219303
[40] Xiuguo, Wu, and Du Shengyong. "An analysis on financial statement fraud detection for Chinese listed companies using deep learning." IEEE Access 10 (2022): 22516-22532. https://doi.org/10.1109/ACCESS.2022.3153478
[41] Li, Shixuan, Wenxuan Shi, Jiancheng Wang, and Heshen Zhou. "A deep learning-based approach to constructing a domain sentiment lexicon: a case study in financial distress prediction." Information Processing & Management 58, no. 5 (2021): 102673. https://doi.org/10.1016/j.ipm.2021.102673
[42] Guo, Changrong, and Jing Xing. "Text Mining-based Enterprise Financial Performance Evaluation in the Context of Enterprise Digital Transformation." International Journal of Advanced Computer Science and Applications 14, no. 6 (2023). https://doi.org/10.14569/IJACSA.2023.01406147
[43] Chi, Yuanying, Mingjian Yan, Yuexia Pang, and Hongbo Lei. "Financial risk assessment of photovoltaic industry listed companies based on text mining." Sustainability 14, no. 19 (2022): 12008. https://doi.org/10.3390/su141912008
[44] Huang, Jian, and Jiangying Wei. "Impact of Intelligent Development on the Total Factor Productivity of Firms-Based on the Evidence from Listed Chinese Manufacturing Firms." Journal of Advanced Computational Intelligence and Intelligent Informatics 26, no. 4 (2022): 555-561. https://doi.org/10.20965/jaciii.2022.p0555
[45] Mao, Jinzhou, Yueyang Zhao, Siying Yang, Rita Yi Man Li, and Jawad Abbas. "Intelligent transformation and customer concentration." Journal of Organizational and End User Computing (JOEUC) 35, no. 2 (2022): 1-15. https://doi.org/10.4018/JOEUC.333470
[46] Olson, David, and Bongsug Chae. "Incorporating an Unsupervised Text Mining Approach into Studying Logistics Risk Management: Insights from Corporate Annual Reports and Topic Modeling." Information 14, no. 7 (2023): 395. https://doi.org/10.3390/info14070395
[47] Kim, Hyun Jung, and Keun Tae Cho. "Analysis of Changes in Innovative Management of Global Insurers in the Preand Post-COVID-19 Eras." Sustainability 14, no. 16 (2022): 9976. https://doi.org/10.3390/su14169976
[48] Yang, Yi, Kunpeng Zhang, and Yangyang Fan. "Analyzing firm reports for volatility prediction: A knowledge-driven text-embedding approach." INFORMS Journal on Computing 34, no. 1 (2022): 522-540. https://doi.org/10.1287/ijoc.2020.1046
[49] Majumdar, Adrija, and Pranav Singh. "Analysis and impact of COVID-19 disclosures: is IT-services different from others?." Industrial Management & Data Systems 123, no. 1 (2023): 345-366. https://doi.org/10.1108/IMDS-04-2021-0239
[50] Tang, Dan, and Jiangying Wei. "Prediction and Characteristic Analysis of Enterprise Digital Transformation Integrating XGBoost and SHAP." Journal of Advanced Computational Intelligence and Intelligent Informatics 27, no. 5 (2023): 780-789. https://doi.org/10.20965/jaciii.2023.p0780
[51] Craja, Patricia, Alisa Kim, and Stefan Lessmann. "Deep learning for detecting financial statement fraud." Decision Support Systems 139 (2020): 113421. https://doi.org/10.1016/j.dss.2020.113421