information

MDPI

Review

Thematic Analysis of Big Data in Financial Institutions Using NLP Techniques with a Cloud Computing Perspective: A Systematic Literature Review

Ratnesh Kumar Sharma [1], Gnana Bharathy [1], Faezeh Karimi [1], Anil V. Mishra [2] and Mukesh Prasad [1,*]

[1] School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney 2007, Australia
[2] School of Business, Western Sydney University, Sydney 2150, Australia
* Correspondence: mukesh.prasad@uts.edu.au

Abstract: This literature review explores the existing work and practices in applying thematic analysis natural language processing techniques to financial data in cloud environments. This work aims to improve two of the five Vs of the big data system. We used the PRISMA approach (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for the review. We analyzed the research papers published over the last 10 years about the topic in question using a keyword-based search and bibliometric analysis. The systematic literature review was conducted in multiple phases, and filters were applied to exclude papers based on the title and abstract initially, then based on the methodology/conclusion, and, finally, after reading the full text. The remaining papers were then considered and are discussed here. We found that automated data discovery methods can be augmented by applying an NLP-based thematic analysis on the financial data in cloud environments. This can help identify the correct classification/categorization and measure data quality for a sentiment analysis.

Keywords: big data; thematic analysis; NLP; finance; thematic metadata; cloud computing; automated data discovery

check for updates

Citation: Sharma, R.K.; Bharathy, G.; Karimi, F.; Mishra, A.V.; Prasad, M. Thematic Analysis of Big Data in Financial Institutions Using NLP Techniques with a Cloud Computing Perspective: A Systematic Literature Review. *Information* **2023**, *14*, 577. https://doi.org/10.3390/info14100577

Academic Editor: Mark Stevenson

Received: 21 August 2023
Revised: 12 September 2023
Accepted: 25 September 2023
Published: 20 October 2023

1. Introduction

A well-designed and -maintained big data platform is an ideal way to assist business users in decision making and provide information access and governance [1–4]. Accurate metadata are crucial for the success of the data lake and enterprise data warehouse environments; however, it is “not” an easy process to achieve the required level of metadata management, information classification, and data security classification and to analyze the shift in customer sentiments over time using manual or semi-automated processes. Generally, the process of defining metadata information is based on the knowledge available to the data owners, data stewards, and data SMEs (subject matter experts), and the process is highly manual and relies on the involvement of a large number of human resources for its end-to-end delivery. Previous studies have shown that metadata information can be generated using automated methods, including NLP-based thematic analyses [1]. The security classification and information classification in the case of large textual data further complicate the creation and maintenance of metadata, and, as per standard practices in large Australian BFSI (banking and financial services institute) organizations, any datasets containing free-form text are considered to contain sensitive/restricted information without further evaluating the information at the record level for any PII (Personally Identifiable Information) [5]. There is a scope to validate the sensitivity of the information at a granular record level (with the use of automated methods) and to make information available to a larger audience.

Continuously changing the data demography with a changing context to improve business processes is a vital source of information for business operational process owners, and a manual time-series analysis is mostly used for this purpose. There is scope to use an NLP-based thematic analysis method to automatically identify the continuously evolving business process improvement areas over a period of time; this can provide a foundation for implementing and achieving trusted, conformed, and highly scalable information assets for up-to-date/reliable information, which can finally help in operational decision making [2], efficiently addressing the issues mentioned above, and improving data lineage.

The main aim of this systematic review is to synthesize the relevant available research on the use of NLP-based thematic analysis methods on financial data using the cloud computing big data environment. This literature review is foundational work for the research topic of metadata management using thematic analyses, the information and security classification of financial data, and sentiment analyses over a period of time using NLP-based methods in the big data cloud environment.

2. Basic Terminology

2.1. Big Data

Large financial organizations across the world started realizing the need to set up big data systems a couple of decades ago, and, since then, the term has become popular [1–11]. Bibri [3] differentiates big data from traditional data technologies in terms of parameters like velocity and variety. There are five Vs of big data.

Volume: Volume matters when it comes to processing and using the information. Volume could range from tens of terabytes to thousands of petabytes. Dealing with the enormous amount of data generation in recent years has become one of the most prominent challenges that organizations face [2,4,6].

Velocity: This is the rate at which datasets are received/ingested and acted on/processed. Mostly, high-velocity data streams write to memory and not to disc for real-time information processing. The five Vs of big data systems are discussed in [4,6].

Variety: This denotes variation in the type of data. Contemporary data formats include semi-structured and unstructured data types, including audio and video, while traditional data formats are strictly relational type.

Veracity: This is the quality and accuracy of the data. The data should be free from impurities and should be in a usable state before being used for making business decisions and operational needs. Veracity is one of the key attributes linked to the "value" of the organizational asset.

Value: This is the actual benefit that big data systems bring to an organization when they are used. It is important to note that the value is directly linked to other attributes like velocity and veracity. Most organizations have the common objective of having massive "Volumes" of high-quality ("Veracity") data of all kinds ("Variety") available at lightning speed ("Velocity") to enterprise systems/users. If any of these attributes are compromised, it directly impacts the success and "Value" that was expected from implementing a big data system in an organization [6].

Using big data in the financial domain and manually tagging data can be a time- and resource-consuming exercise [11]. There has been increased focus on using large text processing in the big data space for many years [1,2,12,13]. This literature review explores the use of an NLP-based thematic analysis on big data in a cloud environment in the financial domain to understand the extent of the research performed in this space.

2.2. Cloud Implementations and Benefits

Cloud environments are preferred platform/infrastructure options because of their flexibility and agility in terms of capacity, expansion, and total cost of ownership (especially during the implementation period); they are being adopted by more and more organizations worldwide, and many of them are cautious about the success of the big data initiatives [6,10].

Cost effectiveness: Cost effectiveness considers the total cost of ownership during the evaluation period. Hariri et al. [4,6,10] wrote that organizations want dynamic data analytics capabilities and to scale the infrastructure as the demands increase. Dynamic provisioning ensures that an adequately sized infrastructure is chosen and that the operational expenses are controlled in line with budgets.

Efficiency to provision: Cloud adoption significantly reduces the planning time needed to set up the infrastructure, and customers only need to make a limited effort in estimating the initial infrastructure and load requirement; this can evolve over a period of time as processing and capacity demand increases over time. This provisioning flexibility dramatically reduces the setup time from months to days. This aspect is of prime importance due to its flexible nature and the growth observed in the big data system [6].

Flexibility: The initial requirement might be to support one of the standard data formats, and the business may soon transform and require integration with other standard live services/applications or may choose to migrate to another cloud provider [6] quickly.

The capability to scale up on demand to process massive volumes in a limited timeframe using multiple parallel processes becomes a pivotal factor in considering cloud solutions in contrast to on-premise infrastructure deployment. Lin et al. [4,6] note that technology owners often complain that the demand has outpaced the change capability when it comes to technology solutions. Cloud adoption enables right-sized environment creation depending on the current requirements of the organization. The infrastructure setup can start small, and more resources can be quickly provisioned on demand. A flexible resource allocation plan works best; e.g. during an uneven demand of resources depending on the load profile, more resources can be added during peak hours of operations instead of letting them be provisioned when there is not much demand. This feature is also called the capability to scale up on demand and caters to the challenges noted by Lin et al. [4,6].

Easy migration to other cloud providers is another benefit and covers the enterprise architecture guidelines of easy transition to another technology platform.

Growth: The growth of cloud adoption between 2014 and 2019 was predicted to increase by 36%, the income prediction was 60% [10], and it was predicted to reach over 3.7 billion people in 2018; there were 7.5% more Internet users than in 2016. Global data production peaked at around 1 zettabyte (ZB) in 2010 and reached 7 ZB by 2014, which became a reason for the increased adoption of cloud big data technologies. Big data, business intelligence (BI), and cloud computing are three possible trends for data monetization due to rising volumes and types of data (see Hanafizadeh [14]).

Popular cloud service providers: The cloud marketplace has risen tremendously over the last decade, resulting in the availability of a broad choice of cloud service providers; however, the market is dominated by three leading players: AWS, Google, and Microsoft Azure. Other cloud vendors like Oracle are slowly catching up. Hariri, Fredericks, and Bowers [10] note that Facebook, Google, and Amazon have each used analytics to harness the potential of big data in their own products. The popularity of Web 2.0 sites like Facebook and Linkedin has significantly boosted data size [15].

2.2.1. Cloud Computing Environments

Depending on the location of the data centers and the interfacing/interaction between them, the cloud computing environments can be mainly categorized as detailed below [16].

Private cloud: A private cloud is one where a dedicated infrastructure is reserved for a particular cloud customer and not shared with other customers. This is an expensive option and is mostly chosen by organizations with extremely high data security and control guidelines/requirements [16].

Public cloud: A public cloud is one where multiple customers share a cloud data center infrastructure (hardware, storage, and network). This is the most cost-effective and commonly used option [16].

Hybrid cloud: A hybrid cloud is one where network interfacing is established between on-premises infrastructure or a private cloud with a public cloud infrastructure [16].

Multi-cloud: A multi-cloud environment is one where more than one cloud platform is used as a disaster recovery primary/secondary option. This is mostly used when disaster recovery requirements are very aggressive and the mean time to recover needs to be kept as low as possible [16].

Regarding extensive data collection and analysis, hybrid, horizontal, and dynamic technologies, like fusing cloud and fog computing, are also worth investigating [7].

2.2.2. Cloud Computing Service Categories

Lin et al. [4] note that cloud services can be segregated into categories such as core infrastructure and networking like storage, computing, and database services. These are used to host application systems, web services, website hosting services, and API services based on the various use cases. Database services fall into relational and non-relational categories. One of the key benefits of cloud database services is that they provide the best in-call information security and maintenance and upgrade infrastructure. Software as a service (SAAS) domain has recently picked up, and services like Google BigQuery and AWS Redshift have become popular options for big data processing and have been adopted by many large organizations [10]. Another key benefit is that SAAS infrastructure saves the cost of employing a full-time niche-skilled workforce to carry out periodic, repetitive application maintenance and administration activities, which are not required as part of the application as a service infrastructure. Capabilities and offerings like IAAS, PAAS, and SAAS work with IAAS as a foundation, and then further features are added in PAAS and SAAS offerings to provide a better experience to customers to save the management and administration effort of COTS applications.

IAAS (infrastructure as a service): This is where only infrastructure components (like servers and storage, networking, and data centers/buildings) are owned by the cloud solution provider. The customer needs to manage the administration and maintenance of the operating environment and tools [4].

PAAS (platform as a service): This is where, in addition to infrastructure components, the operating systems and the development environment are owned by the cloud solution provider. The customer must manage the administration and maintenance of the core application hosted on the cloud platform [4].

SAAS (software as a service): This is where the cloud solution provider manages everything. The customer manages only the operational administration of the core application hosted on the cloud platform [4].

2.3. Thematic Analysis

A thematic analysis is a qualitative analysis method applied over a set of text and involves coding documents and identifying themes [17,18]. Codes are the foundational element for generating themes, and this captures the qualitative attributes of the data relevant to the subject/aim of the analysis [19]. The thematic analysis approach proposed by Broun and Clarke is the most popular. Figure 1 shows the various steps involved in a thematic analysis. The first step is the familiarization step, where an overall understanding is achieved, and the relation to the objective of the research topic is ascertained. In further steps, documenting and coding are carried out, which can be considered as the tokenization of the data. Irrelevant words (stop words, etc.) are removed to clean up the data and retain the relevant information only, which can help generate meaningful themes. Finally, all the coding outputs are observed to generate themes by looking at the patterns in the codes and by checking the correlation between them. The generation of themes using coded data can be challenging and depends on factors like familiarity with the overall domain and the subject of the research. It is essential to associate the meaning of codes with the domain of the research. That is, words related to the financial domain need to be considered; for instance, stress would point to financial stress and would not be related to the medical field. Once the initial themes are determined, then these are reviewed in multiple iterations to finalize the themes and conclude the results [17,20]. Refer to Figure 1 for the steps involved in the thematic analysis process.

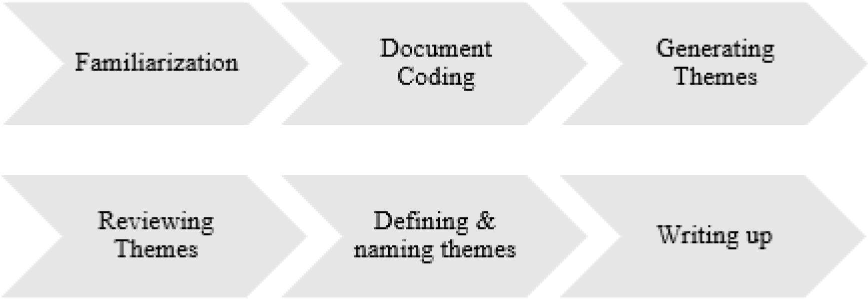

Figure 1. Steps and flow in thematic analysis.

2.4. Financial Data Categories

Organizational data are a complex subject, and correctly organizing data is one of the most challenging exercises for an organization. Advanced analytic techniques have always been a tool to find underlying information, uncovering any patterns, and to establish correlations between very large data components [10,14]. Lima et al. [2] discuss mapping/associating the data with the correct classification category and its associated benefits in big data environments. Depending on the definition and operational boundaries of the enterprise, data town planning is the first step, which provides a basic framework to define data subject areas. Then, entities can be classified and associated with these subject areas after conducting a deep data analysis to automatically associate or classify the entities, especially when these are textual objects in a non-relational format. The data town plan lays the foundation and provides a blueprint of the connection between various subject areas, e.g., assets, interactions, opportunities, products, parties, customers, employee, claims, service requests, and financial and non-financial transactions. The organizational data town plan information and data objects can be mapped after the thematic analysis of the information content of the data objects. Figure 2 contains various categories in which financial data are organized in the financial domain.

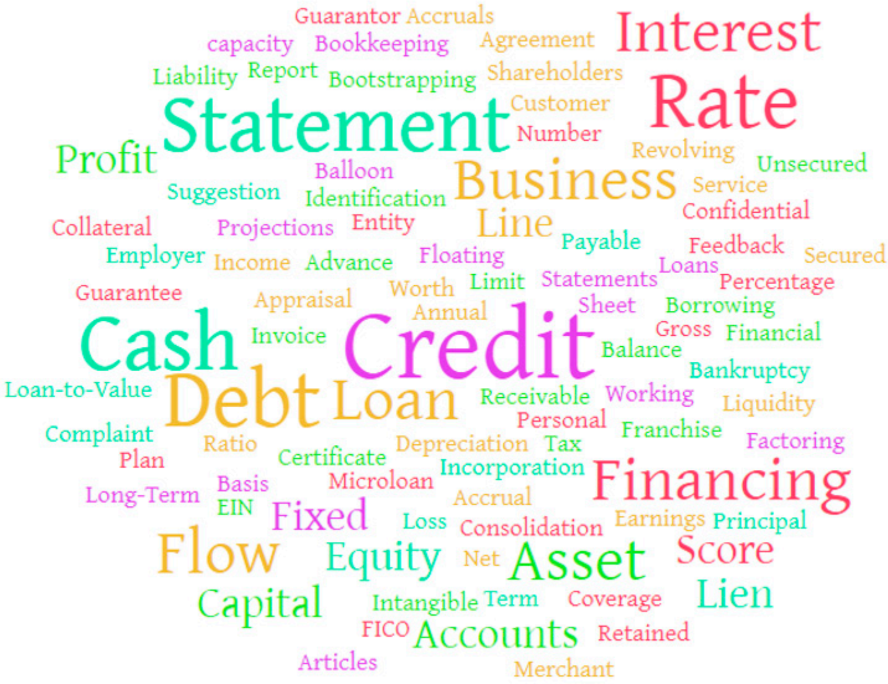

Figure 2. Financial data categories.

3. Literature Review

A thorough literature review is carried out to assess the depth of the research work conducted so far in the subject mentioned above in phases. The PRISMA literature review methodology is followed as part of this work. Our study analyzes the research papers published over the last 10 years about the topic in question using a keyword-based search and bibliometric analysis. The systematic literature review is conducted in multiple phases, and filters are applied to exclude papers based on title and abstract initially, then based on the methodology/conclusion, and, finally, after reading the full text. The remaining papers are considered and discussed here. Table 1 shows the various steps followed in the literature review.

Table 1. Stages of the literature review conducted using the PRISMA technique.

Stage	Description
Stage 1	Identifying keywords and databases and performing searches based on individual keywords. n = ~1825 for bigdata finance n = ~693 for nlp finance
Stage 2	Assessing the results involving the PRISMA approach. Searching based on a combination of keywords. n = 73(nlp, bigdata, finance) + 73(ThematicAnalysis, nlp) + 63(ThematicAnalysis, bigdata) = 209
Stage 3	Excluding studies based on title and abstract n = 123
Stage 4	Excluding studies based on abstract/methodology/conclusion n = 98
Stage 5	Including studies after reading the full text and discussion of results n = 53

A list of keywords is prepared, and based on all combinations of the keyword families, a search is performed on the database using individual keywords (n = ~1825 for bigdata finance, and n = ~693 for nlp finance).

The search criteria are further refined, and a search is performed using a combination of keywords (n = 73 (nlp, bigdata, finance) + 73 (ThematicAnalysis, nlp) + 63 (ThematicAnalysis, bigdata) = 209). The results are reviewed independently to avoid any possibility of bias in line with the PRISMA technique for systematic literature reviews. The list is initially narrowed down based on the title and abstract (n = 123) and then further narrowed down based on the abstract, methodology, and conclusion (n = 98).

For the remaining papers, the full text is read before concluding whether they are relevant to the research objective (n = 53). Peer reviews and inputs are taken in line with the PRISMA technique.

Thematic analysis techniques have been prominently used in medical research, as evidenced by the number of papers published [21,22].

Many other researchers rely on social media platforms for the availability of research data and use various methods for data sourcing [21]. Some people use tools like "culturintel" to source social media data [23] for research purposes, while others use tools like decahose [22]. Overall, the availability of social media data makes such a data a well-known alternative to using confidential financial data. Hariri et al. [10] talk about the issues of noise, incompleteness, and inconsistency in data. Mondal [24] advocates the use of NLP for data standardization. Lima et al. [2] use Knime and weka for sourcing and Hbase to store news data. Van, Le-Khac, and Kechadi [25] compare the performance of a popular financial investigation tool (FTK) with that of a custom-developed application, LES, but they do not discuss the architectural differences and internal workings of the tools. FTKs' push-down optimization support for big data is also not explored; however, it is noted that thematic analysis has been the prevalent approach for data segregation and classification. Latent Dirichlet Allocation (LDA) is found to be the most popular method for topic modeling [22], while at the same time, some authors choose to select users from different groups for a similar exercise [26]. In thematic analyses of social media data, the use of Python LDA techniques and visualization using popular tools is commonly observed; however, the rare use of a Python-based visualization library called “pyldv” by Zheng et al. has also been observed [27], while some researchers use R library-based approaches like STM (structural topic modeling) [28].

Perez [29] applied transformer-based topic modeling to conduct a thematic analysis. Similarly, computer-assisted qualitative data analysis software (CAQDAS), such as MAXQDA and NVivo, also provide some degree of NLP-enabled thematic analysis, as noted by Chang et al. [30]. Andreotta et al. [31] suggest an iterative approach to topic grouping by repeatedly using batches of 5, 10, and 20 topics. Akter et al. [32] use a survey data sample. Hanafizadeh and Harati [14] conduct a literature review using the thematic analysis technique. Lin et al. [4] discuss four prominent themes after their thematic analysis of survey data, which are data revolution, high technology, transformation, and strategy, and they suggest that these are focus areas for a modern enterprise. However, Bibri [3,7,33] discusses the use of big data and related technologies in smart cities. Akter et al. [34] propose a six-step mechanism to properly execute BDA projects to support data-driven decision making for service organizations. Pedro and Hart [6] discuss the need for a big data system; however, they do not discuss the need for suitable metadata, which are critical for the success of the big data initiatives of an enterprise. Repeated application of the classification algorithm, as noted by Hariri et al. [10], could prove to be beneficial in improving the accuracy of metadata for effective information classification.

Pedro, Brown, and Hart [6] discuss the strategic vision and alignment with big data practice; however they do not mention the need for the availability of information metadata assets. In contrast, Bibri and Anshari et al. [3,5,7,8,33] emphasize the need to create context-aware applications coupled with big data technologies, and they note that real-time capabilities can provide a better grip on the process flow for decision making. Che, Zhu, and Li [35] discuss conducting a three-dimensional sentiment analysis to relate the CRS activities of an organization and its financial performance to impact investors’ confidence. The three perspectives from which the analysis is performed are dictionary (micro-level), theme (meso-level), and machine learning (macro-level) perspectives. Mbah, Rege, and Misra [36] advocate the use of spark and scala over a live stream of data, focus on the LSA approach, and discuss considering the essential concepts/terms, but they do not consider each and every word/term used in the document as an approach to deal with the noise by utilizing the singular value decomposition (SVD) method. Bhardwaj et al. [12] present a primary method for predicting stock market performance based on the index values. The paper contains a good description of various machine learning and prediction/classification methods. Gu, Harkoff et al. [37,38] suggest that domain-specific context is crucial for text mining and producing compelling insights. The authors advocate the use of BIM (business intelligence model), a method similar to UML; however, it is essential to note that such modeling applications are scarce, especially in modern banking enterprises, which are mainly moving away from such traditional practices and embracing agile methodologies.

Hujala, Klein et al. [39,40] advocate the use of automated methods to process large amounts of text, and they note that, even after applying unsupervised methods for data processing, including LDA, there is still a need for human intervention to validate and keep the output meaningful and relevant to the context. Odlum [41] uses the (HierNMF2) method to generate tree nodes for topic modeling. Tang et al. [42] found that the BERT method for intention classification worked better than the TextCNN, HAN, and ELMO methods when they used it for transfer learning use cases on insurance data. Ni et al. [43] use individual sentiment score computation with a fixed set of rules to derive public opinion about stocks to categorize posts as negative/positive under bullish/bearish criteria. Li et al. [9] find the term vector method to be a widely used method for representing text in a computer-friendly form, while Chan et al. [1] compare multiple methods like GI and sentiwordnet, run a sentiment analysis engine (SAE) with the random and bag-of-words approach, and note considerable improvements. O’Halloran [11] claims to have discovered hidden patterns after applying computational data science techniques that are not recognized by coding rules. The method provides improvement over manual coding methods; however, it still relies on conventional data seeding and training methods.

Chen et al. [44] propose a tool to model public mood and emotions, which is mainly based on the information available in popular news channels and financial blobs; however, there are chances that this is biased and the subject of market manipulation since it does not account for the voices of real investors from primary/secondary markets, the majority of whom make use of popular social media platforms like Twitter and Facebook and may not necessarily write blogs. The authors, however, made good use of TFIDF and lexicon training to generate the model. Esichaikul and Phumdontree [45] apply models like CNN, Dynamic seq2-CNN, CNN-Bidirectional GRU, Basic Dynamic CNN, and CNN-Bidirectional LSTM to perform a sentiment analysis on financial data from Thailand's websites. They find CNN-Bidirectional GRU to be the most effective method for this analysis. Yan et al. [46] match lenders with borrowers and find that the random forest algorithm works better than the gradient boosting method, while Tsaih et al. [13] use aspects of speech tagging along with other methods like word tokenization on time-series data and finally create a machine learning model to predict the currency exchange rate/forex. Konstantinidis et al. [47] study the impact of social media news/sentiments on asset prices and market movement, while Ma R. et al. (2017) perform a comparative analysis.

While NLP-based techniques enable high efficiency and scalability, they have also received criticism, often from practitioners who take the traditional approach. Skeen et al. [48] point out that NLP techniques produce ″coarse descriptions″ that lack the nuances that a manual approach could bring to the table.

With the advent of large language models (LLMs), there is potential for augmenting the thematic analysis's accuracy, efficiency, and nuances. Sallam demonstrates and reviews the use of ChatGPT in health research tasks. However, researchers such as Sallam and Watkins warn that the use of LLMs should be approached with care, owing to misinterpretations, bias and risk of misinformation, data privacy and cybersecurity issues, plagiarism, and a lack of attributions, among other reasons. These researchers rightly warn that there is a need to develop appropriate guidelines for the use of LLMs in research [49–51].

3.1. Bibliometric Analysis

3.1.1. Annual Scientific Production

We analyzed the number of research papers published yearly using the results obtained after the keyword search. As shown in Figure 3, there was a significant dip in the number of articles published in the year 2016, after which the publication of articles gained momentum and continuously increased ever since (as evident by the dashed trendline). There was a significant increase in the number of papers published afterward (solid line showing graph data points).

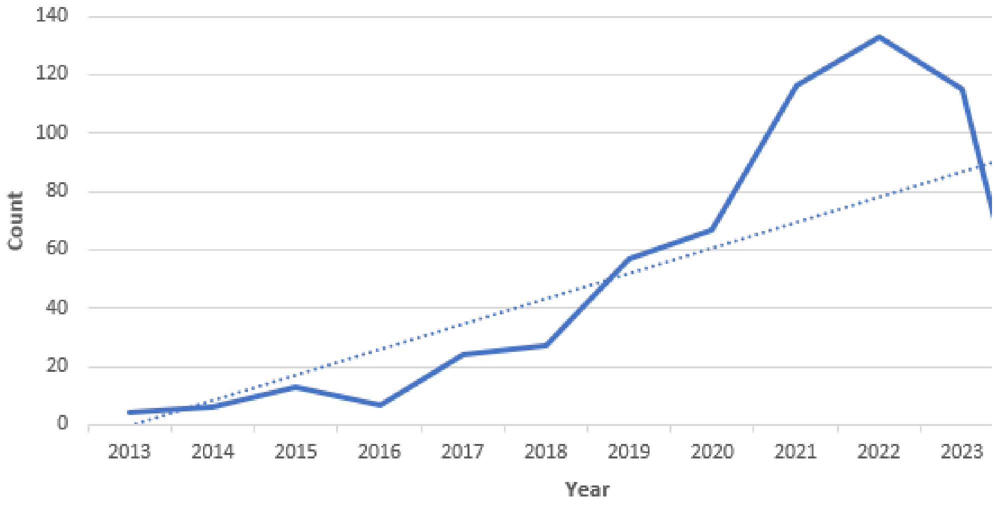

Figure 3. Bibliometric analysis—annual counts of scientific production of research papers.

3.1.2. Citation Analysis

We analyzed the number of article citations per year using the results obtained after the keyword search. As shown in Figure 4, a spike in the number of citations was observed between the years 2017 and 2019. In contrast, the dip in citations during the year 2016 shows that there needed to be a stronger focus on this research subject during this time. It was also observed that, between the years 2017 and 2018, the citation counts almost plateaued (solid line showing graph data points). However, from an overall trend perspective, the research keywords have been a focus area of research work, as evidenced by the overall increase in the number of article citations (as shown by the dashed trendline).

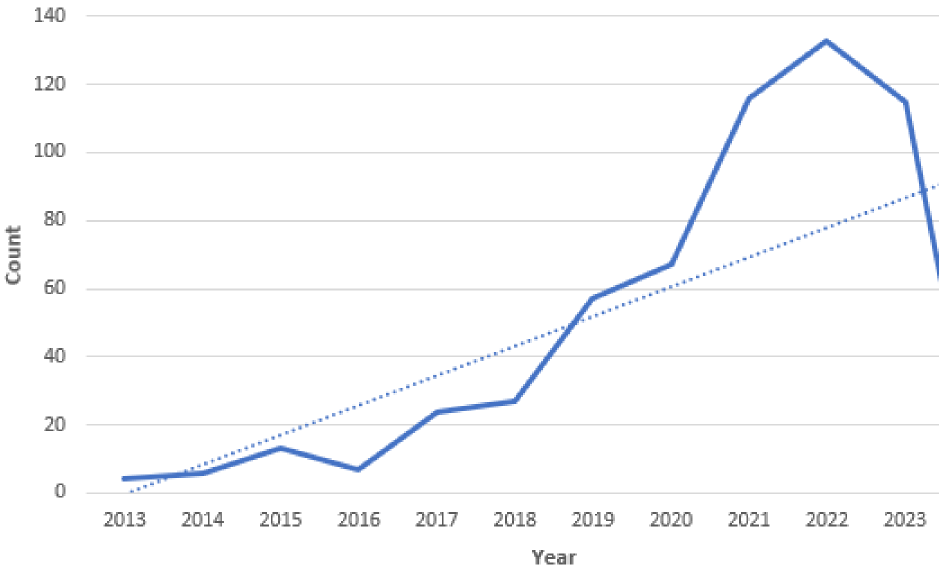

Figure 4. Bibliometric analysis—research citations by year.

3.1.3. Key Stats

Table 2 contains key information regarding the research documents related to the topic published between the years 2013 and 2022. Our study found around 142 documents from around 116 sources and 573 authors. According to the author statistics observed, most of the authors published more than one paper, except for 10 authors, who published a single paper, resulting in a total of 615 appearances of all authors. The limited number of documents published by single authors indicates a good amount of collaboration between the authors in this research area. The average number of citations per document is above 5 for this time period; however, the average number of citations per year is around 1.5. A variety of document types was observed, with the majority of document types being research articles and conference papers.

Table 2. Key stats (literature review).

	Description	Results
Main Information about Data	Timespan	2013:2022
	Sources (journals, books, etc.)	116
	Documents	142
	Average years from publication	2.01
	Average citations per document	5.352
	Average citations per year per doc	1.524
	References	1

Table 2. *Cont.*

	Description	Results
Document Types	Article	91
	Book chapter	1
	Conference paper	35
	Conference review	1
	Letter	3
	Review	11
Authors	Authors	573
	Author appearances	615
	Authors of single-authored documents	10
	Authors of multi-authored documents	563
Author Collaboration	Single-authored documents	10
	Documents per author	0.248
	Authors per document	4.04
	Co-authors per document	4.33
	Collaboration index	4.27

Table 3 contains the number of published documents from various journals/sources. It needs to be noted that the field of medical research is the most popular domain where big data and thematic analysis techniques are used and research documentation is published.

Table 3. Bibliometric analysis—most relevant sources ranked based on number of articles produced.

Journal Name	No. of Articles Produced
Journal of Medical Intern	10
Lecture Notes in Computer	6
JMR Formative Research	3
JMR Public Health and Su	3
PLOS One	3
BMJ Open	2
CEUR Workshop Proceedings	2
Clinical Toxicology	2
Journal of Advanced Research	2
Journal of Hospitality and	2
Lecture Notes in Electric	2
Others	1

3.1.4. Co-Occurrence Network

The network map in Figure 5 shows the relation between the keywords related to the research. Multiple keyword groups/clusters are identified and shown in different colors. Figure 5 shows various keywords related to one family, e.g., medical-domain-related words like COVID (or other variances) and pandemic, forming a cluster. However, they are sized according to the number of occurrences and grouped and clustered with other words depending on how frequently they are grouped together in a single research paper.

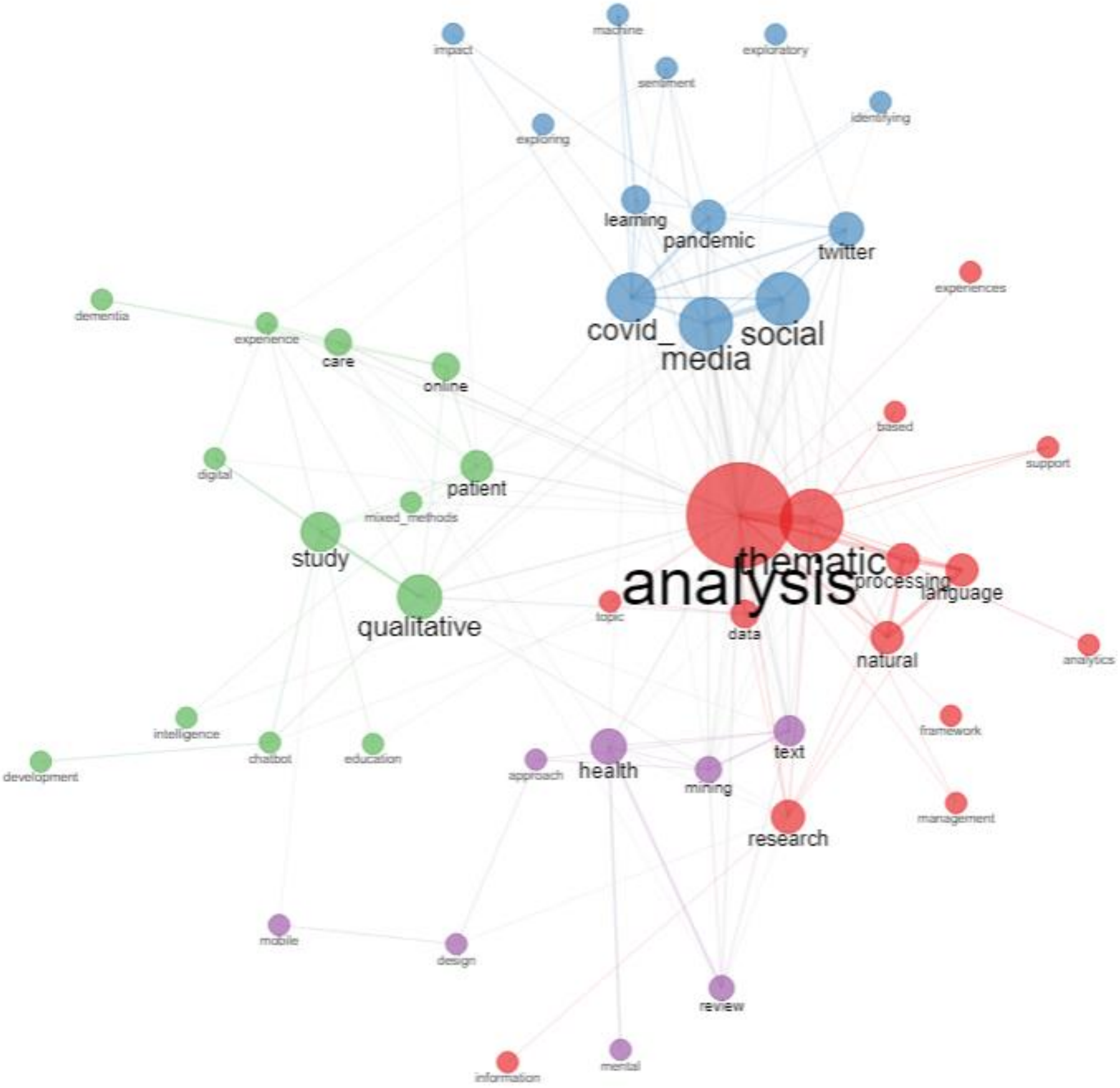

Figure 5. Bibliometric analysis—co-occurrence network of relevant keywords.

4. Discussion

Modern financial organizations aspire to create data assets quickly and efficiently with the use of automated methods for information onboarding, information classification, and data security classification and to further analyze the shift in customer sentiments [1] over a period of time in order to provide a desired level of customer experience and remain competent in the market. The findings suggest that a thematic data analysis using natural language processing (NLP) techniques in the financial domain can be applied to the data in the cloud to classify metadata, classify information in line with the information security guidelines, or determine the main themes (e.g., customer interactions like feedback/complaints [5]) and how they change over a period of time; these are some of the very innovative use cases that can help financial organizations achieve the goals mentioned above and to improve their operational effectiveness.

Based on the findings, it can be concluded that the efficient use of enterprise information assets to serve customers better is the top goal of today's modern enterprises. Such enterprises believe in the continuous improvement of organizational processes [4,6,24,52] to serve the customer, which would result in overall improved business outcomes, improved return on investment (ROI) [16], and reduced total cost of ownership (TCO) [53]. TCO is defined as the combined cost of capital expenditure and operational expenditure over a period of time for comparison purposes. In other words, this is the complete cost of setting up and running the infrastructure for a period of time. ROI is a critical success factor for any investment made by a business and is considered while comparing the overall cost to the organization versus the value that it brings in. It is important to note that ROI is not visible during the early lifecycle of the adoption of any cloud service, and an accurate picture only becomes visible after a considerable period of time. If the benefits outweigh the investment, it is considered a positive return on investment.

A big data ecosystem should not only be considered as an enterprise data repository but should also be coupled and have a direct interface with customer-serving context-aware applications to maximize its value. The correct metadata of the information content is a crucial driver for efficient information sharing and analysis. It is critical to have automated controls and methods available to traverse the data over a period of time, to ascertain the correctness of the information, and to correctly categorize the information based on the study of customer sentiments. This could help in analyzing the deficiencies in business operations and the information catalog, allowing for corrections to be made in a timely manner to ensure that the enterprise data are up to date and relevant for business users [4]. In this literature review, gaps were identified regarding the use of NLP-based thematic analysis methods for the discovery of metadata and for determining their relationship with financial data town plan information so that data objects can be correctly mapped. The two related aspects are metadata discovery and information content classification and their correlation with the financial data town plan. This can potentially improve the veracity of the big data environment and, ultimately, its value.

No studies were found regarding the use of NLP techniques in the big data domain for the classification of sensitive information/information security to help information governance for financial organizations. The increasing trend in research citations based on the research keywords proves that this is a popular research topic; however, no relevant results were found when using the combination of all research keywords, thus confirming the existence of a research gap. There is no evidence of any study having been conducted to analyze the changes in data demography and customer sentiments over a period of time. This could have helped pinpoint a business process that needs to be in focus during a certain period of time, which is linked to the overall satisfaction/dissatisfaction of the consumer/customer. This literature review suggests that there is scope for bridging this gap with further research on the application of thematic analysis techniques on financial data in the big data cloud computing domain, which makes this a novel research topic.

5. Conclusions and Future Work

5.1. Conclusions

A systematic literature review is conducted using the PRISMA technique and a bibliometric analysis to synthesize the available research on the use of NLP-based thematic analysis methods on financial data using the cloud computing big data environment. This review suggests that the adoption of big data analytics and cloud computing has been expanding rapidly and notably enhances the capabilities of the industry. Though NLP approaches have recently been utilized more frequently, the use cases are still evolving, and they are most commonly used for sentiment analyses according to the existing studies.

This literature review acknowledges a gap in the use of NLP-based thematic analysis methods to create accurate metadata, information content classification for information security requirement classification, and for sentiment analyses over a period of time. Using these methods will enable financial organizations to gauge the accuracy of and maintain big data assets and provide the right information so that this organizational asset can be used correctly and efficiently to achieve customer satisfaction goals, even when the profile of the data changes rapidly.

Thematic analysis techniques have been mostly used in the field of medical research. The findings are consistent with those of previous research, confirming that thematic analysis techniques are popular for literature review purposes and for sentiment analyses in the medical research domain. These methods can potentially be applied to generate/discover metadata information for big data/data lake environments, which could improve the veracity and, ultimately, the value of the big data platform for the financial organization.

5.2. Future Work

There are further aspects and dimensions that can be explored to enhance the usability of enterprise information by creating the right type of information channels/interfaces to ensure that this information consumed or added to by IoT devices and sensors, and the information content (sentiment) and any other visual indicator may be correlated in a timely manner for the business/customer. Further research can explore the use of this information to assess/predict the customer's mood based on their previous interactions and current facial expressions by making use of all this relevant information from the big data/data lake environment and other interaction channels used.

Author Contributions: Conceptualization, R.K.S. and M.P.; methodology, R.K.S. and F.K.; software, R.K.S. and G.B.; validation, G.B., F.K., A.V.M. and M.P.; formal analysis, R.K.S. and A.V.M.; investigation, R.K.S. and G.B.; resources, G.B., F.K., A.V.M. and M.P.; data curation, R.K.S.; writing—original draft preparation, R.K.S.; writing—review and editing, R.K.S., G.B., F.K., A.V.M. and M.P.; visualization, R.K.S. and F.K.; supervision, M.P.; project administration, G.B. and F.K.; funding acquisition, A.V.M. and M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chan, S.W.K.; Chong, M.W.C. Sentiment analysis in financial texts. *Decis. Support. Syst.* **2017**, *94*, 53–64. [CrossRef]
2. Lima, L.; Portela, F.; Santos, M.F.; Abelha, A.; Machado, J. Big data for stock market by means of mining techniques. In *Advances in Intelligent Systems and Computing*; Springer: Berlin/Heidelberg, Germany, 2015.
3. Bibri, S.E. The anatomy of the data-driven smart sustainable city: Instrumentation, datafication, computerization and related applications. *J. Big Data* **2019**, *6*, 59. [CrossRef]
4. Lin, C.; Kunnathur, A.S.; Li, L. Conceptualizing big data practices. *Int. J. Account. Inf. Manag.* **2020**, *28*, 205–222. [CrossRef]
5. Anshari, M.; Almunawar, M.N.; Lim, S.A.; Al-Mudimigh, A. Customer relationship management and big data enabled: Personalization & customization of services. *Appl. Comput. Inf.* **2019**, *15*, 94–101.
6. Pedro, J.; Brown, I.; Hart, M. Capabilities and Readiness for Big Data Analytics. *Proc. Comput. Sci.* **2019**, *164*, 3–10. [CrossRef]
7. Bibri, S.E. Sustainable Urban Forms: Time to Smarten up with Big Data Analytics and Context–Aware Computing for Sustainability. In *Smart Sustainable Cities of the Future*; Elsevier: Amsterdam, The Netherlands, 2018.
8. Bibri, S.E.; Krogstie, J. ICT of the new wave of computing for sustainable urban forms: Their big data and context-aware augmented typologies and design concepts. *Sustain. Cities Soc.* **2017**, *32*, 449–474. [CrossRef]
9. Li, Q.; Chen, Y.; Wang, J.; Chen, Y.; Chen, H. Web Media and Stock Markets: A Survey and Future Directions from a Big Data Perspective. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 381–399. [CrossRef]
10. Hariri, R.H.; Fredericks, E.M.; Bowers, K.M. Uncertainty in big data analytics: Survey, opportunities, and challenges. *J. Big Data* **2019**, *6*, 44. [CrossRef]
11. O'Halloran, S.; Maskey, S.; McAllister, G.; Park, D.K.; Chen, K. Big data and the regulation of financial markets. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 25–28 August 2015.
12. Bhardwaj, A.; Narayan, Y.; Dutta, M. Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. *Proc. Comput. Sci.* **2015**, *70*, 85–91. [CrossRef]
13. Tsaih, R.H.; Kuo, B.S.; Lin, T.H.; Hsu, C.C. The Use of Big Data Analytics to Predict the Foreign Exchange Rate Based on Public Media: A Machine-Learning Experiment. *IT Prof.* **2018**, *20*, 34–41. [CrossRef]

14. Hanafizadeh, P.; Harati Nik, M.R. Configuration of Data Monetization: A Review of Literature with Thematic Analysis. *Glob. J. Flex. Syst. Manag.* **2020**, *21*, 17–34. [CrossRef]
15. Li, R.Y.M.; Song, L.; Li, B.; James, C.; Crabbe, M.; Yue, X.G. Predicting Carpark Prices Indices in Hong Kong Using AutoML. CMES Comput. *Model. Eng. Sci.* **2023**, *134*, 2247–2282.
16. Arunachalam, D.; Kumar, N.; Kawalek, J.P. Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. *Transp. Res. Part. E Logist. Transp. Rev.* **2018**, *114*, 416–436. [CrossRef]
17. Clarke, V.; Braun, V. Thematic analysis. *J. Posit. Psychol.* **2017**, *12*, 297–298. [CrossRef]
18. Braun, V.; Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
19. Boyatzis, R. *Transforming Qualitative Information: Thematic Analysis and Code Development*; Sage: Thousand Oaks, CA, USA, 1998.
20. Braun, V.; Clarke, V. *Successful Qualitative Research: A Practical Guide for Beginners. Successful Qualitative Research: A Practical Guide for Beginners*; Sage: Thousand Oaks, CA, USA, 2013.
21. Petersen, K.; Gerken, J.M. #COVID-19: An exploratory investigation of hashtag usage on Twitter. *Health Policy* **2021**, *125*, 541–547.
22. Xiang, X.; Lu, X.; Halavanau, A.; Xue, J.; Sun, Y.; Lai, P.H.L.; Wu, Z. Modern Senicide in the Face of a Pandemic: An Examination of Public Discourse and Sentiment About Older Adults and COVID-19 Using Machine Learning. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **2021**, *76*, e190–e200. [CrossRef]
23. Falcone, T.; Dagar, A.; Castilla-Puentes, R.C.; Anand, A.; Brethenoux, C.; Valleta, L.G.; Furey, P.; Timmons-Mitchell, J.; Pestana-Knight, E. Digital conversations about suicide among teenagers and adults with epilepsy: A big-data, machine learning analysis. *Epilepsia* **2020**, *61*, 951–958. [CrossRef]
24. Mondal, B. Artificial intelligence: State of the art. In *Book Artificial Intelligence: State of the Art*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 389–425.
25. Van Banerveld, M.; Le-Khac, N.A.; Kechadi, M.T. Performance evaluation of a natural language processing approach applied in white collar crime investigation. In *Future Data and Security Engineering*; Springer: Berlin/Heidelberg, Germany, 2014.
26. Guntuku, S.C.; Schneider, R.; Pelullo, A.; Young, J.; Wong, V.; Ungar, L.; Polsky, D.; Volpp, K.G.; Merchant, R. Studying expressions of loneliness in individuals using twitter: An observational study. *BMJ Open* **2019**, *9*, e030355. [CrossRef]
27. Zheng, C.; Xue, J.; Sun, Y.; Zhu, T. Public opinions and concerns regarding the Canadian prime minister's daily COVID-19 briefing: Longitudinal study of youtube comments using machine learning techniques. *J. Med. Internet* **2021**, *23*, e23957. [CrossRef]
28. Rodriguez, M.Y.; Storer, H. A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data. *J. Technol. Hum. Serv.* **2020**, *38*, 54–86. [CrossRef]
29. Pérez, V.; Caro, R.; Rua Vieites, A. Unraveling the Complexities of Climate Change and Environment Migration: A Transformers-Based Topic Modelling Approach; 2023, preprint version. [CrossRef]
30. Chang, T.; DeJonckheere, M.; Vydiswaran, V.G.V.; Li, J.; Buis, L.R.; Guetterman, T.C. Accelerating Mixed Methods Research With Natural Language Processing of Big Text Data. *J. Mix. Methods Res.* **2021**, *15*, 398–412. [CrossRef]
31. Andreotta, M.; Nugroho, R.; Hurlstone, M.J.; Boschetti, F.; Farrell, S.; Walker, I.; Paris, C. Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis. *Behav. Res. Method.* **2019**, *51*, 1766–1781. [CrossRef] [PubMed]
32. Akter, S.; Gunasekaran, A.; Wamba, S.F.; Babu, M.M.; Hani, U. Reshaping competitive advantages with analytics capabilities in service systems. *Technol. Forecast. Soc. Chang.* **2020**, *159*, 120180. [CrossRef]
33. Bibri, S.E. The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability. *Sustain. Cities Soc.* **2018**, *38*, 230–253. [CrossRef]
34. Akter, S.; Bandara, R.; Hani, U.; Fosso Wamba, S.; Foropon, C.; Papadopoulos, T. Analytics-based decision-making for service systems: A qualitative study and agenda for future research. *Int. J. Inf. Manag.* **2019**, *48*, 85–95. [CrossRef]
35. Che, S.; Zhu, W.; Li, X. Anticipating Corporate Financial Performance from CEO Letters Utilizing Sentiment Analysis. *Math. Probl. Eng.* **2020**, *4*, 5609272. [CrossRef]
36. Mbah, R.B.K.; Rege, M.; Misra, B. Using spark and scala for discovering latent trends in job markets. In Proceedings of the ICCDA 2019: Proceedings of the 2019 3rd International Conference on Compute and Data Analysis, New York, NY, USA, 14–17 March 2019.
37. Gu, Y.; Storey, V.C.; Woo, C.C. *Conceptual Modeling for Financial Investment with Text Mining*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2015.
38. Horkoff, J.; Barone, D.; Jiang, L.; Yu, E.; Amyot, D.; Borgida, A.; Mylopoulos, J. Strategic business modeling: Representation and reasoning. *Softw. Syst. Model.* **2014**, *13*, 1015–1041. [CrossRef]
39. Hujala, M.; Knutas, A.; Hynninen, T.; Arminen, H. Improving the quality of teaching by utilising written student feedback: A streamlined process. *Comput. Educ.* **2020**, *157*, 103965. [CrossRef]
40. Klein, L.F.; Eisenstein, J.; Sun, I. Exploratory thematic analysis for digitized archival collections. *Digit. Scholarsh. Humanit.* **2015**, *30*, i130–i141. [CrossRef]
41. Odlum, M.; Yoon, S.; Broadwell, P.; Brewer, R.; Kuang, D. How twitter can support the HIV/AIDS response to achieve the 2030 eradication goal: In-depth thematic analysis of world AIDS day tweets. *JMIR Public Health Surv.* **2018**, *4*, e10262. [CrossRef] [PubMed]

42. Tang, S.; Liu, Q.; Tan, W.A. *Intention Classification based on Transfer Learning: A Case Study on Insurance Data*; Springer: Berlin/Heidelberg, Germany, 2019.
43. Ni, Y.; Su, Z.; Wang, W.; Ying, Y. A novel stock evaluation index based on public opinion analysis. *Proc. Comput. Sci.* **2019**, *147*, 581–587. [CrossRef]
44. Chen, M.Y.; Chen, T.H. Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena. Future Gen. *Comput. Syst.* **2019**, *96*, 692–699. [CrossRef]
45. Esichaikul, V.; Phumdontree, C. Sentiment analysis of Thai financial news. In Proceedings of the ICSEB'18: Proceedings of the 2018 2nd International Conference on Software and e-Business, New York, NY, USA, 18–20 December 2018.
46. Yan, J.; Wang, K.; Liu, Y.; Xu, K.; Kang, L.; Chen, X.; Zhu, H. Mining social lending motivations for loan project recommendations. *Expert. Syst. Appl.* **2018**, *111*, 100–106. [CrossRef]
47. Konstantinidis, A.; Scalzodees, B.; Calvi, G.G.; Mandic, D.P. *Text Mining—A Key Lynchpin in the Investment Process: A Survey*; Series Frontiers in Artificial Intelligence and Applications, Applications of Intelligent Systems; IOS Press: Amsterdam, The Netherlands, 2018; Volume 310, pp. 181–193. [CrossRef]
48. Skeen, S.; Jones, S.; Cruse, C.; Horvath, K. Integrating Natural Language Processing and Interpretive Thematic Analyses to Gain Human-Centered Design Insights on HIV Mobile Health: Proof-of-Concept Analysis. *JMIR Hum. Factors* **2022**, *9*, e37350. [CrossRef]
49. Sallam, M. ChatGPT Utility in Health Care Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef]
50. Watkins, R. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI Ethics* **2023**, 6–7. [CrossRef]
51. Sallam, M. The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations. *medRxiv* **2023**. [CrossRef]
52. Yang, N. Financial Big Data Management and Control and Artificial Intelligence Analysis Method Based on Data Mining Technology. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 7596094. [CrossRef]
53. Suciu, G.; Suciu, V.; Halunga, S.; Fratu, O. Big data, internet of things and cloud convergence for E-Health applications. In *Book Big Data, Internet of Things and Cloud Convergence for E-Health Applications*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 151–160.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.