electronics

MDPI

Article

Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study

Weisi Chen [1,*], Fethi Rabhi [2], Wenqi Liao [1] and Islam Al-Qudah [3]

[1] School of Software Engineering, Xiamen University of Technology, Xiamen 361024, China; 2212114218@s.xmut.edu.cn
[2] School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia; f.rabhi@unsw.edu.au
[3] Faculty of Computer Information Science, Higher Colleges of Technology, Abu Dhabi P.O. Box 25026, United Arab Emirates; ialqudah@hct.ac.ae
* Correspondence: chenweisi@xmut.edu.cn

Abstract: News impact analysis has become a common task conducted by finance researchers, which involves reading and selecting news articles based on themes and sentiments, pairing news events and relevant stocks, and measuring the impact of selected news on stock prices. To facilitate more efficient news selection, topic modeling can be applied to generate topics out of a large number of news documents. However, there is very limited existing literature comparing topic models in the context of finance-related news impact analysis. In this paper, we compare three state-of-the-art topic models, namely Latent Dirichlet allocation (LDA), Top2Vec, and BERTopic, in a defined scenario of news impact analysis on financial markets, where 38,240 news articles with an average length of 590 words are analyzed. A service-oriented framework for news impact analysis called "News Impact Analysis" (NIA) is advocated to leverage multiple topic models and provide an automated and seamless news impact analysis process for finance researchers. Experimental results have shown that BERTopic performed best in this scenario, with minimal data preprocessing, the highest coherence score, the best interpretability, and reasonable computing time. In addition, a finance researcher was able to conduct the entire news impact analysis process, which validated the feasibility and usability of the NIA framework.

Keywords: topic modeling; news analysis; finance; LDA; Top2Vec; BERTopic

Citation: Chen, W.; Rabhi, F.; Liao, W.; Al-Qudah, I. Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study. *Electronics* **2023**, *12*, 2605. https://doi.org/10.3390/electronics12122605

Academic Editor: Domenico Ursino

Received: 5 May 2023
Revised: 6 June 2023
Accepted: 7 June 2023
Published: 9 June 2023

1. Introduction

1.1. Background and Motivation

In the era of big data, the expansion of news media channels in recent years has accelerated the dissemination of news data. Since then, news impact analysis has become one of the main tasks carried out by scholars and practitioners in a wide range of studies. In general, the analysis methods employed in this type of research entail deciphering the content of news articles before determining their impact on a particular field. Many methods of news impact analysis have been proposed and adopted, including sentiment analysis or opinion mining [1], which determines the sentiment represented in the text by calculating sentiment scores [2,3]; and text mining [4], which is used to detect patterns in text and discover new insights. These methods are part of the natural language processing (NLP) family [5].

The analysis and interpretation of how financial markets respond to or behave in reaction to news is extremely complicated. Research on analyzing the impact of news on financial markets has been conducted in two ways. Some studies concentrate on modeling and evaluating financial markets from the perspective of finance specialists without applying advanced techniques, e.g., [6–8], whereas others build and validate novel text mining or opinion mining techniques from a technical perspective, without paying

sufficient attention to domain-specific impact analysis models, e.g., [9–11]. To bridge the gap between these two classes of studies, in our preliminary work [12] we have advocated a systematic service-oriented framework to facilitate reproducible news sentiment impact analysis processes in the context of financial markets, enabling finance domain experts to evaluate the impact of any news dataset on the related financial market instruments. The news impact analysis process, as shown in Figure 1, includes setting financial context parameters, importing news data, selecting news based on the topics and sentiments, pairing news events and financial instruments (e.g., stocks and cryptocurrencies [13]), importing financial market data, and performing impact analysis and visualization. One gap in the old framework version is in the news selection stage. It has embedded merely lexicon-based sentiment analysis to identify news articles with extreme emotions. However, with a vast number of news articles, finance domain experts still need to scan or read through the content of each news article, to determine what topic it belongs to and what financial instrument it relates to before performing the impact analysis, which is time-consuming and unrealistic, especially for large-scale news impact analysis.

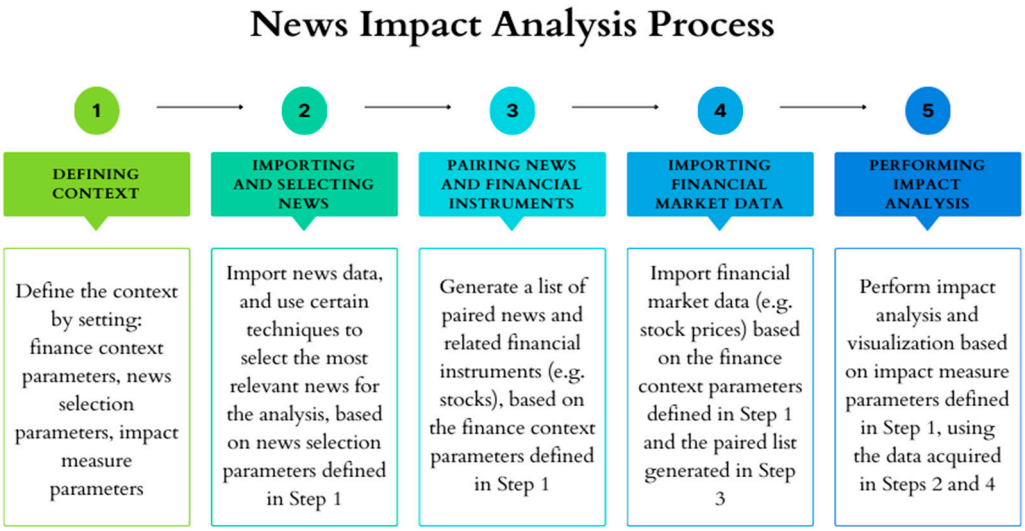

Figure 1. Existing news impact analysis process.

Topic modeling (TM) [14] has sprung up as a subset of NLP, aiming to extract topics or themes from a collection of texts or documents, also known as corpus, such as books, websites, blogs, social media postings, emails, news items, research articles, etc. It is one of the most important text-mining techniques that falls under the unsupervised learning umbrella. Compared with supervised learning methods, where texts are labeled correctly, topic modeling does not require manual labeling prior to the analysis, resulting in higher efficiency in figuring out the clustered topics within texts. Yet, the current reality is that most finance researchers still need to manually read through news items, discover topics of interest, and select the most relevant news, which is time-consuming, and they usually find it difficult to opt for and operate the most appropriate TM technique, as machine learning knowledge and programming skills are often involved.

1.2. Research Question and Contributions

This article attempts to address two unanswered research questions:

RQ 1: How do state-of-the-art TM techniques compare in the financial news context, especially with large-scale, long news corpus?

Due to the barrier of technology and the wide variety of TM techniques, finance domain experts either have to manually discover topics via a "read-and-decide" approach or rely on an IT or NLP expert to select and implement certain TM techniques for specific research tasks. There are currently no guidelines on which state-of-the-art TM technique may perform better in the financial news context. Therefore, addressing this research

question is critical to bridge this gap. In this article, we examine large-scale, long finance-related news articles and discover topics using a number of the most prominent TM techniques, and compare the results based on coherence, interpretability, and computation time, to answer this research question.

RQ 2: How can topic modeling be incorporated into the financial news impact analysis framework? What is the extended process?

In many financial research tasks, finance domain experts usually assume that certain events may have a positive or negative impact on the financial market, e.g., stocks related to the events, and attempt to measure how these events would affect the stock prices. In practice, finance experts usually discover topics out of news articles manually and decide which news items are of interest in their research tasks accordingly. The topic discovery step considers both the theme and the sentiment of each news item, aiming to select the most relevant news. After that, the finance expert would measure how the financial market, e.g., stocks related to the selected news, react to the occurrences of these news articles. However, there is currently no systematic way of leveraging state-of-the-art TM techniques for news impact analysis in the existing literature. Thus, this research question is of great importance for applied NLP in the finance domain.

In this paper, we provide a comparison among various state-of-the-art TM techniques in the context of financial news impact analysis, aiming to assist finance domain experts in making informed decisions when conducting such tasks, and meanwhile extending our formerly advocated framework by enabling the TM capabilities. To the best of our knowledge, there is very limited existing literature on applying and comparing various most advanced TM techniques, and our work is the first of this kind in the context of large-scale and long finance-related news, which would benefit both finance researchers and NLP practitioners who conduct finance-related news analysis tasks. Specifically, the contributions of this article include the following:

- Reviewing and summarizing state-of-the-art TM techniques.
- Integrating TM techniques with the existing financial news impact analysis process.
- Comparing and evaluating TM techniques in the context of financial news impact analysis with large-scale and long financial news corpus.
- Providing a systematic method and guidelines for finance domain experts and NLP practitioners to conduct financial news impact analysis leveraging TM capabilities.

1.3. Structure of the Article

The rest of the paper is organized as follows. Section 2 reviews the various most up-to-date text modeling techniques and their applications. Section 3 describes the extended News Impact Analysis (NIA) framework designed to facilitate reproducible news impact analysis with any news dataset, as well as the incorporated topic modeling techniques. Section 4 demonstrates the experiment conducted to compare the performance of selected state-of-the-art topic modeling techniques and validate the proposed framework, followed by Section 5 concluding the paper, raising limitations of the study, and highlighting future research directions.

2. Literature Review

Topic modeling (TM) is an unsupervised learning method in natural language processing (NLP) to cluster text-based documents according to the latent semantic structure. It has been studied for several decades, and many methods have been proposed. From the perspective of underlying algorithms, topic modeling techniques can be categorized into three major types, namely algebraic, probabilistic, and neural models [15,16]. The first two are conventional statistics-based methods, and the last one is the more up-to-date neural artificial neural network-based technique since the proliferation of applying deep learning to NLP. Figure 2 describes these categories along with key models that belong to each category.

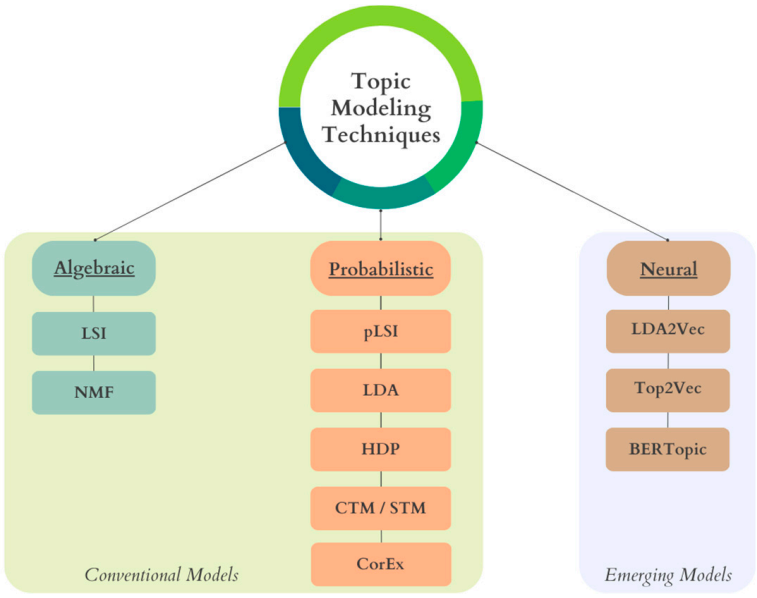

Figure 2. Categories of topic models.

Algebraic models include latent semantic indexing (LSI), non-negative matrix factorization (NMF), and probabilistic models, including probabilistic latent semantic indexing (pLSI), anchored correlation explanation (CorEx) [17], latent Dirichlet allocation (LDA) [18], and many LDA extensions and variants such as hierarchical Dirichlet process (HDP), correlated topic model (CTM), and structural topic model (STM). It should be noted that STM is created expressly with social science research in mind, which enables the inclusion of metadata into the model and reveals how different texts may discuss the same fundamental topic in diverse ways. Among these aforementioned conventional models, LDA has been the most commonly used method for decades since its inception [18]. However, most studies of this kind have utilized LDA for granted, most likely due to its popularity, and failed to provide the reasoning for the topic modeling method selection. LDA can be considered a generalized version of pLSI by including a Dirichlet prior distribution across the document-topic and topic-word distributions. One limitation of LDA is that it adopts the bag-of-words (BoW) method to represent texts, which essentially disregards semantics between words within the text. In general, conventional TM techniques such as LDA require fairly complex corpus pre-processing; careful selection of parameters, such as the number of topics to be generated; appropriate model evaluation; and interpreting the generated topics based on common sense and domain knowledge.

Most recently developed neural models have emerged and gained increasing popularity since 2016. Examples of the neural category include lda2vec [19], SBM [20], deep LDA [21], Top2Vec [22], and BERTopic [23]. The development trace is in line with the exponential advancement of deep learning during the last few years. In particular, the older deep LDA is a hybrid model combining LDA and a basic multilayer perceptron (MLP) neural network. In contrast, the most recent BERTopic is based on the more advanced bidirectional encoder representations from transformer (BERT) and class-based TF-IDF (c-TF-IDF). It has become increasingly dominant in the topic modeling field due to its demonstrated capabilities of producing cutting-edge outcomes on a number of datasets with minimal data preprocessing.

Over the past decade, TM has been actively utilized in various domains, such as health, hospitality, education, social networks [24], and finance, benefiting both academia and industry, especially interdisciplinary studies. Recent literature shows that most studies focus on applying particular TM techniques (mostly one technique) to a specific domain. Table 1 summarizes some applied research employing TM techniques identified in recent literature.

Table 1. Ten applied research studies using topic modeling in recent literature.

Ref.	Year	Corpus and Size	Length	TM Technique	Details
[25]	2018	Webpages (10,000)	long	LDA	Applying LDA to identify websites related to food safety issues and highlighting the potential of LDA as a valuable tool for communication researchers.
[26]	2019	Research papers (650)	long	LDA	Smart literature review conducted by loading research papers, using LDA to generate topics, and selecting appropriate topics for literature review.
[27]	2019	Hotel reviews (27,864)	long	STM	Analyzing New York City hotel reviews using STM and showing that it improves inferences about consumer dissatisfaction.
[28]	2020	Research papers (3963)	long	STM	Utilizing STM to identify research topics out of the title, keywords, and the abstract of articles published in the journal Computers & Education over 42 years.
[29]	2021	News (100,000)	long	Top2Vec	Identifying the most widely reported topics or issues within COVID-related news in outlets of UK, India, Japan, and South Korea using Top2Vec, followed by news sentiment analysis using the RoBERTa model.
[30]	2021	Forum posts and Tweets (Unknown size)	short	LDA	Analyzing bitcoin-related posts on Twitter, Reddit and Bitcoin Talk using LDA, and the result is then used by an LSTM-based neural system for stock price prediction.
[31]	2022	Tweets (31,800)	short	LDA NMF Top2Vec BERTopic	Adopting and comparing four TM techniques on social media data for the purpose of social science research. NMF and BERTopic performed better than the other two in this scenario.
[32]	2022	Tweets (78,827)	short	LDA	Extracting the topics using LDA and sentiment polarity using a dictionary-based sentiment analysis method out of the tweets related to the COVID-19 vaccine.
[33]	2022	Instagram (33,881)	short	LDA CorEx NMF	Utilizing three topic modeling techniques to identify traveler experiences out of Instagram posts with a certain hashtag in 2020.
[34]	2023	News (2158)	long	LDA	Topic modeling on financial news using LDA, and highlighting predictions and speculative statements within text via a graphical user interface.

LDA has been the most frequently seen TM technique in the recent literature, e.g., [25,26,30,32,34]. These studies all demonstrate the possibility of using TM to discover topics in various fields. There have been very few studies leveraging TM on finance-related tasks. A recent study has used LDA to discover topics from short Tweets and generates predictions on financial instruments [34]. To the best of our knowledge, currently, there is a research gap on utilizing and comparing multiple state-of-the-art TM techniques on long finance-related news, to facilitate the analysis of news impact on the financial market following research needs in finance, which leaves the two research questions defined in Section 1.2 unanswered.

3. News Impact Analysis Using Topic Modeling Techniques

To facilitate the analysis of news impact on the financial market with topic modeling capabilities following research needs in finance, a News Impact Analysis (NIA) framework is proposed. This framework is the key to answering RQ 2 as identified in Section 1.2, i.e., to provide a systematic method to leverage state-of-the-art TM techniques for the news impact analysis process. This section will start by introducing the NIA framework that our work is based on from the software engineering perspective, and then elaborate on the TM techniques selected for comparison purposes in this article.

3.1. NIA Framework

The NIA framework extends our preliminary work [12], which comprises three elements; a Parameters Model (PM), a software architecture, and a defined process. The defined process of performing financial news impact analysis has been illustrated in Figure 1.

The PM is composed of two types of parameters, which jointly define the context of a given news impact analysis task, namely finance context parameters (FCP) and news selection parameters (NSP). Details of these parameters are shown in Figure 3.

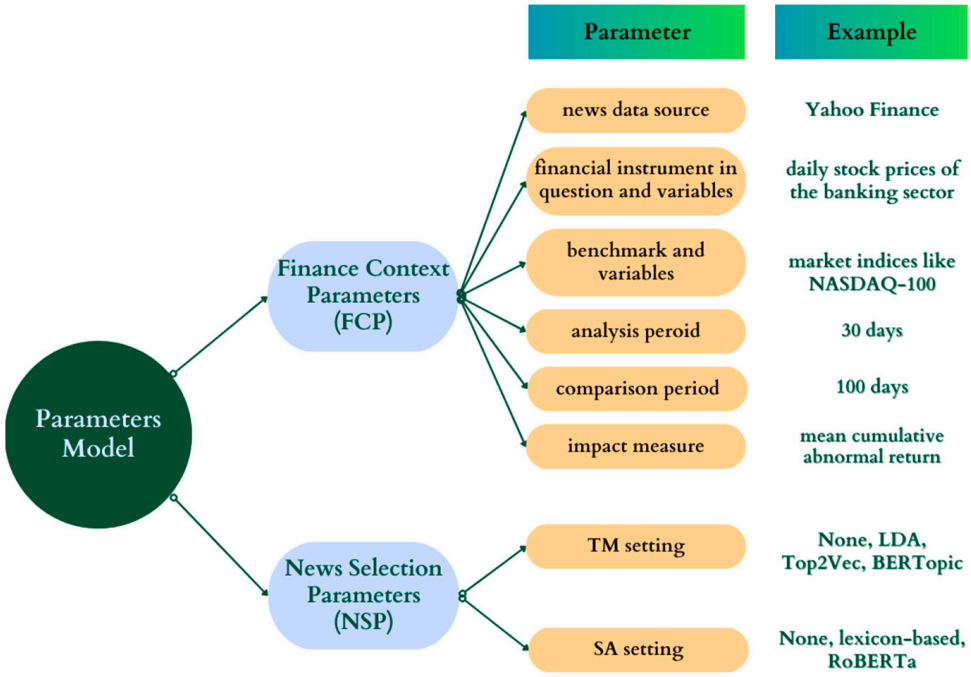

Figure 3. NIA Parameters Model (PM).

Most importantly, NSP includes specific sub-level parameters for topic modeling and sentiment analysis, depending on the model selection. For instance, if LDA is selected as the TM method, the number of topics should be specified; if a lexicon-based technique is selected as the SA method, the lexicon and the threshold of the sentiment score should be defined.

The NIA software architecture offers guidelines for implementing news impact analysis. It follows a service-oriented architecture (SOA), which consists of three distinct layers, namely the user layer, the service layer, and the data layer, as shown in Figure 4. The user layer provides the user interface for interacting with the service layer, allowing end-users to define the PM parameters and invoke services. The service layer contains services that encapsulate the business logic of the analysis process, while the data layer stores all datasets generated by the services in the service layer. In this proposed extension, the user can opt to use only one of the Topic Modeling Service and the Sentiment Analysis Service to filter the news by topic or sentiment, or both in one study. Table 2 describes the core services and their interactions within this framework.

3.2. Integrated Topic Modeling Techniques

In this study, we have incorporated three topic modeling techniques into the Topic Modeling Service of the NIA framework for comparison purposes, including one most prevalent conventional model (LDA) and two emerging deep learning-based models (Top2Vec and BERTopic). The details of these models are described as follows. The implementation details of these models will be described in Section 4.1.2.

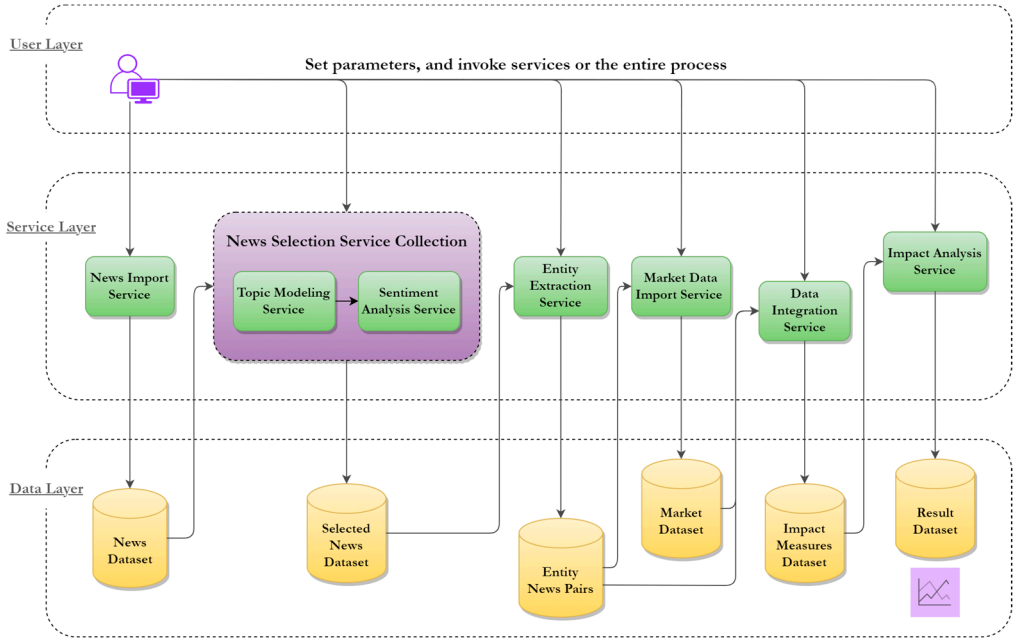

Figure 4. Extended software architecture for news impact analysis.

Table 2. Details of core services in the NIA framework.

Service Name	Description
News Import Service	Responsible for importing news data from a variety of data sources as per the user-defined PM and feeding it into the Data Layer as the News Dataset.
Topic Modeling Service	Conducting topic modeling on the News Dataset, filtering the news by topic based on the user-defined PM, and then generating the Selected News Dataset in the Data Layer, or feeding the result into the following Sentiment Analysis Service for further news filtering.
Sentiment Analysis Service	Generating sentiment scores and identifying extreme news based on the user-defined PM. Results are committed to the Selected News Dataset in the Data Layer.
Entity Extraction Service	Generating the Entity-News Pairs in the Data Layer and updates the PM accordingly, based on user selections or the content of selected news in the Selected News Dataset.
Market Data Import Service	Responsible for importing financial market data based on the defined PM and the Entity-News Pairs. The data are saved as the Market Dataset in the Data Layer.
Data Integration Service	Merging the Entity-News Pairs and the Market Dataset into the Impact Measures Dataset in the Data Layer.
Impact Analysis Service	Performing impact analysis based on the PM. Results that are committed to the Results Dataset in the Data Layer along with some data visualization.

3.2.1. LDA

LDA [18] is a well-known generative probabilistic model for identifying topic information latent in a large document collection or corpus. The approach is predicated on the assumption that each document represents a probability distribution composed of a number of topics, and each topic represents a probability distribution composed of many words. It uses the bag-of-words (BoW) approach, which treats each document as a vector of word frequencies, thus transforming textual information into numerical information that can be easily modeled. Essentially, LDA reduces the dimensionality of the bag-of-words model by representing a document as a topic. The number of topics is usually a few hundred, representing the document as a vector of a few hundred dimensions, greatly speeding up training and making it relatively less prone to overfitting.

LDA defines a generative process including: (1) a topic is drawn from the topic distribution for each document; (2) extracting a word from the word distribution corresponding to the extracted topic; (3) repeating the process until every word in the document is traversed. More formally, each document in the corpus corresponds to a multinomial distribution of T topics, denoted as θ. Each topic corresponds to a multinomial distribution of words in the vocabulary, denoted as φ. The vocabulary comprises all the mutually exclusive words in all the documents in the corpus, but some stop words must be excluded, and some stemming must be performed in the actual modeling. θ and φ have a Dirichlet prior distribution with hyperparameters α and β, respectively. For each word in document D, a topic z from the multinomial distribution θ corresponding to that document is extracted. Then a word w from the multinomial distribution φ corresponding to topic z is extracted. This process is repeated N times to produce document D, where N is the total number of words in document D. This process is shown in Figure 5.

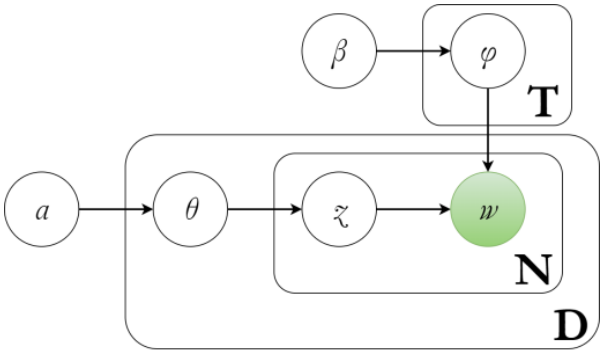

Figure 5. The defined generative process of LDA.

3.2.2. Top2Vec

Top2Vec [22] is an unsupervised machine learning approach developed to provide scalable and effective topic modeling and document clustering solutions. To discover the most relevant themes in large-scale text corpora, this method leverages a hierarchical clustering algorithm that utilizes word embedding semantic similarity to organize the documents into coherent clusters, which are then assigned to the most representative topics. Specifically, the method maps both documents and words to a common semantic vector space using the Doc2Vec method. The document vectors are then clustered into several clusters, each representing a distinct topic. The topic representation of a given cluster is derived by averaging the document vectors within the cluster and extracting the N nearest words to the topic vector. Notably, Top2Vec does not require prior knowledge of the number of topics and can handle multi-word phrases and infrequently used terms, setting it apart from traditional topic modeling techniques.

Top2Vec comprises a series of steps. First, embedding vectors and words are generated (commonly using Doc2Vec). Next, the dimensionality of the embedding vectors is reduced (commonly by using UMAP). Subsequently, clustering is performed on the reduced vectors (commonly via HDBSCAN). Afterward, the centroids of the resulting clusters are computed, each representing a distinct topic. The vector of each topic is obtained by averaging all the document vectors within the same cluster. Finally, topic assignment is carried out by associating words that are in close proximity to the vector of each cluster.

3.2.3. BERTopic

BERTopic [23] is a cutting-edge pre-trained topic modeling technique that leverages BERT and c-TF-IDF to construct dense clusters that facilitate the interpretation of topics while retaining significant words in topic descriptions. Unlike conventional topic modeling techniques, BERTopic employs the powerful contextualized word embeddings provided by BERT to capture the semantics and context of words in a corpus. Moreover, BERTopic fea-

tures a user-friendly interface that allows researchers to observe and analyze the outcomes of the topic modeling process.

Akin to Top2Vec, BERTopic involves embedding documents, reducing dimensionality (using UMAP), clustering (using HDBSCAN), and generating topic representations from clusters. It is worth noting that the final step entails utilizing c-TF-IDF to extract topic words and decrease the number of topics and apply maximum marginal relevance (MMR) to improve word coherence and diversity. The processes of both Top2Vec and BERTopic are depicted in Figure 6. Note that one key factor that distinguishes Top2Vec and BERTopic from LDA is that Top2Vec and BERTopic feature continuous topic modeling, whereas LDA provides discrete modeling [35].

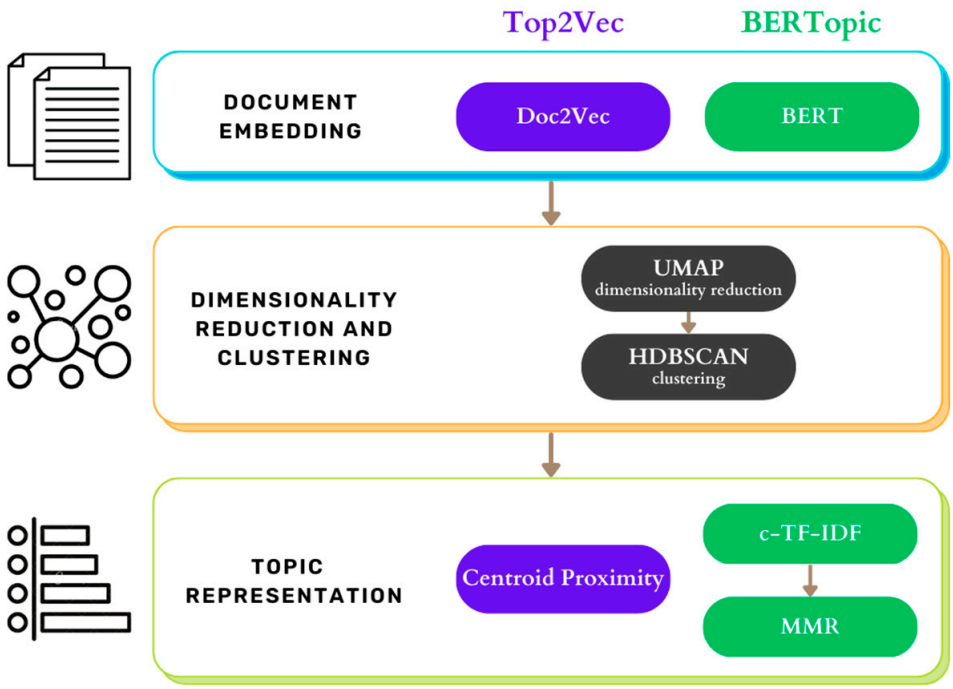

Figure 6. The generative process of Top2Vec and BERTopic.

4. Experiments and Results

This section documents the experiments conducted for this study and demonstrates the results in detail. These experiments aim to compare various TM results in the context of a defined finance research scenario, the result of which will answer RQ 1 as identified in Section 1.2; and to validate the NIA framework proposed in Section 3, which answers RQ 2 as identified in Section 1.2.

4.1. Experimental Setup

4.1.1. Dataset

In this experiment, we gained access to news data from an Australian mainstream newspaper called the Australian Financial Review (AFR) in XML format, with a total corpus size of 981 MB (219,538 news articles in English). We wrote a Python program, embedded in the News Import Service, to extract 38,240 news articles in the "Companies and Markets" category from 1 January 2015 to 31 December 2021, with an average length of 29.12 sentences and 590.14 words. Figure 7 displays a sample of the news dataset, and we used the full news text in the "text" field for analysis purposes.

	headlines	bylines	text
0	Pressure on Morrison for fire inquiry	Tom McIlroy	Scott Morrison is under growing pressure to la...
1	RBA's 2020 test on interest rates	Sarah Turner, Vesna Poljak and Robert Guy	The Reserve Bank is expected to slash interest...
2	Small-town Canada to BHP boss	Timothy Moore and James Thomson	Abbotsford, British Columbia | Modest. It's th...
3	Treasurer warned on risky funds	John Kehoe	Dozens of listed investment funds that have ra...
4	Toll rising in fire crisis	Bo Seo, James Fernyhough and Tom McIlroy	The death toll may exceed eight people as the ...
...	...	...	...
1566	Gladesville, Sydney $2.1 million	NaN	Lawyer Ellen Knoblanche and her husband, Allan...
1567	Pavilion style	NaN	Hawthorn East, Melbourne $3.9 million to 4.2 ...
1568	Dress refresh	John Davidson	A closet that's full of hot air.Ick. I've been...
1569	Balmoral, Brisbane $2.5 million	NaN	Zac Krstev and David Liekari of development co...
1570	TIME OUT	Life & Leisure	James Ciuffetelli executive general manager Ye...

Figure 7. A sample of the news data used in the experiment.

4.1.2. Hardware and Software Prototype Implementation

The Topic Modeling Service has been implemented using Python, which has integrated three TM techniques, namely LDA, Top2Vec, and BERTopic, using Python libraries scikit-learn (Version 1.1.1), top2vec (Version 1.0.29), and BERTopic (Version 0.14.1), respectively. Table 3 provides implementation details of these topic models, including required Python libraries and the parameters used in this study. The Sentiment Analysis Service has been implemented using Python integrating a RoBERTa-based model [36]. The rest of the services have been implemented using the R language. All services have their RESTful API exposed so they can be called by other services and the user interface. The prototype has been used for experiments in this study.

Table 3. Details of topic model implementation in this study.

Item	LDA	Top2Vec	BERTopic
Python library	scikit-learn	top2vec	bertopic
version	1.1.1	1.0.29	0.14.1
No. of topics	20, 50, 100, 500	undefined	undefined
max iterations	1, 10	undefined	undefined
min topic size	undefined	undefined	10
dimensionality reduction	undefined	UMAP	UMAP
clustering	undefined	HDBSCAN	HDBSCAN
topic representation	default	centroid proximity	c-TF-IDF, MMR

The hardware where the experiments were run is as follows:

- Operating System: Windows 11 64 bits
- CPU: 11th Gen Intel(R) Core (TM) i7-11370H @ 3.30 GHz
- RAM: 16 GB

4.1.3. Scenario

The defined scenario is that a finance researcher would like to discover how news concerning the banking sector affects the Big Four banks in Australia. The researcher first needs to discover the relevant bank-related news using the Topic Modeling Service, filter the selected news further by sentiment using the Sentiment Analysis Service, and then generate a list of news-entity pairs, including the news dates, which will be used as event dates (Day 0) of the impact analysis. Finally, the impact analysis process will measure how significant the events (news) would result in abnormal returns (the difference between the actual return of a stock and the expected return as per the benchmark) of the big four banks in Australia. The NIA parameters used in this case study are illustrated in Table 4.

Table 4. NIA parameters used in the experiment.

Parameter Name	Parameter Value
News data source	News with the type of "Companies and Markets", sourced from AFR
Financial instruments	Daily close prices of the stocks of the Big Four Australian banks (security codes: CBA, WBC, NAB, and ANZ), sourced from Yahoo Finance
Benchmark	All Ordinaries index, sourced from Yahoo Finance
TM setting	LDA (with 20, 50, 100, 500 topics & 1 and 10 max iterations), Top2Vec, and BERTopic
SA setting	RoBERTa (threshold = "mean sentiment score")
Analysis period	(−20 days, +20 days)
Comparison period	(−100 days, −21 days)
Impact measure	Mean cumulative abnormal returns (MCAR)

In this experiment, after importing the news data via the News Import Service, the Topic Modeling Service was executed ten times to facilitate our comparison between various topic models, including LDA (with eight different sets of parameters), Top2Vec, and BERTopic; and then the finance researcher used the "bank-related" news items out of the best-performing models to run through the entire news impact analysis process to validate the extended framework described in Section 3.1, including invoking the Sentiment Analysis Service to further select negative news items out of the ones selected by the Topic Modeling Service, creating a list of news-entity pairs using the Entity Extraction Service, importing relevant stock market data using the Market Data Import Service, merging all the required datasets using the Data Integration Service, generating the impact analysis results via the Impact Analysis Service, and finally interpreting the impact analysis results to see if it makes sense from the financial perspective.

4.2. Results

4.2.1. LDA Results

We have trained the LDA model using eight different sets of parameters, with a combination of different numbers of topics to be generated (20, 50, 100, and 500) and different numbers of maximum iterations (1 and 10). We have recorded the computation time and the coherence scores for all topics generated by each run. The finance researcher inspected the topics and keywords, tried to interpret the results, and provided annotations of a sample of topics. The annotations indicate the possible links between the generated topics and various societal aspects or industrial sectors. Tables 5 and 6 show a sample of 10 topics generated by the LDA-20-10 model (20 topics and 10 maximum iterations) and the LDA-100-10 model (100 topics and 10 maximum iterations), respectively. Some of the topic keywords did not make much sense to the finance researcher, which were marked as blanks in the tables. Figure 8 presents a visualization of the generated topics by the LDA-100-10 model, indicating the statistical proximity of topics.

Table 5. Sample topics generated by LDA 20-10 (20 topics and 10 maximum iterations).

No.	Top-10 Keywords	Annotation
Topic 1	financial, commission, report, court, regulator, claim, review, action, government, case	economy
Topic 2	bank, loan, credit, financial, banking, capital, billion, customer, rate, risk	banking
Topic 3	rate, economy, global, china, economic, bond, world, central, investor, policy	economy
Topic 4	project, group, construction, contract, infrastructure, road, toll, building, billion, contractor	construction
Topic 5	share, shareholder, board, deal, group, investor, offer, capital, executive, director	investment
Topic 6	crown, network, medium, telstra, mobile, casino, nbn, news, service, content	communication
Topic 7	share, price, growth, profit, earnings, month, billion, stock, analyst, result	investment
Topic 8	say, people, time, executive, chief, big, like, make, think, way	-
Topic 9	energy, power, solar, electricity, vehicle, car, renewable, battery, generation, wind	energy
Topic 10	coal, port, rail, union, queensland, worker, thermal, aurizon, terminal, agreement	energy

Table 6. Sample topics generated by LDA 100-10 (100 topics and 10 maximum iterations).

No.	Top-10 Keywords	Annotation
Topic 1	deal, billion, merger, agreement, asset, acquisition, transaction, talk, potential, offer	merger
Topic 2	davis, healy, cromwell, quicksilver, pierre, deleted, familiarity, inman, lewinns, callaghan	-
Topic 3	bain, cochlear, hearing, tasmanian, implant, piper, fish, remark, salmon, private	hearing
Topic 4	santos, pipeline, gas, cooper, apa, williams, narrabri, central, coates, mccormack	energy
Topic 5	woolworth, food, coles, supermarket, sale, sup0plier, price, chain, product, retailer	retail (grocery)
Topic 6	fuel, caltex, refinery, petrol, arrium, whyalla, refining, viva, conversion, ampol	petrol
Topic 7	bhp, bhps, mackenzie, billion, shale, henry, exploration, production, scarborough, asset	mining
Topic 8	coal, thermal, tonne, queensland, coking, miner, hunter, mining, export, whitehaven	mining
Topic 9	store, retailer, retail, sale, online, brand, myer, chain, customer, jones	retail
Topic 10	bank, banking, customer, westpac, anz, nab, cba, commonwealth, royal, banker	banking

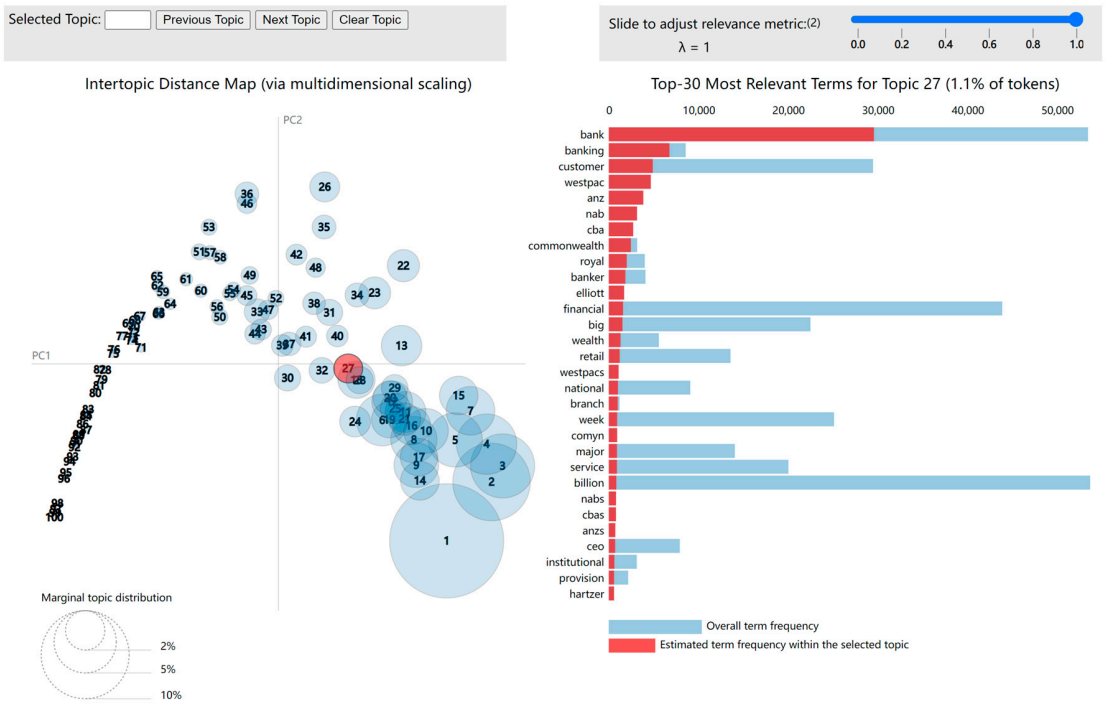

Figure 8. Visual representation of LDA-generated topics (100 topics and 10 maximum iterations).

4.2.2. Top2Vec Results

The Top2Vec model automatically determines the number of topics, and it does not require complex data preprocessing (e.g., eliminating stop words), as seen in LDA. We have trained the Top2Vec model using its default settings. Similarly, the finance researcher inspected, interpreted, and annotated a sample of the generated topics. Table 7 shows a sample of 10 topics. Compared with the LDA models in Section 4.2.1, the topics generated by Top2Vec were mostly interpretable. As an example, Figure 9 presents a word cloud based on a particular topic related to the term "banking". The news items that belong to such topics may then be used for impact analysis on the banking sector.

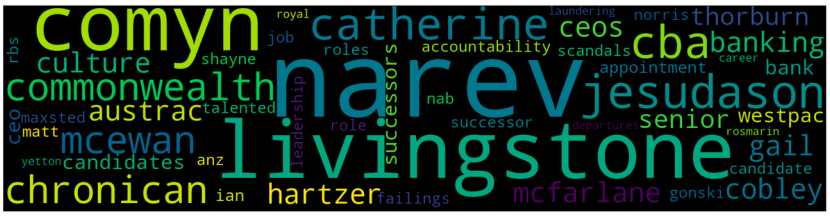

Figure 9. Word cloud for one of the Top2Vec-generated topics based on "banking".

Table 7. Sample topics generated by Top2Vec.

No.	Top-10 Keywords	Annotation
Topic 1	steadyoil, steady, hang, seng, shanghai, pm, changecash, nikkei, commodities, yr	commodities
Topic 2	economists, rba, inflation, reserve, economist, recession, dales, hike, dovish, unemployment	economy
Topic 3	emissions, climate, carbon, greenhouse, fossil, decarbonisation, warming, emitting, emission, decarbonise	climate
Topic 4	fundie, stocks, conviction, caps, managers, stockpicker, quant, investing, you, equities	investment
Topic 5	coles, supermarket, grocery, supermarkets, banducci, woolworths, cain, aldi, groceries, durkan	retail (grocery)
Topic 6	strategists, stocks, strategist, cassidy, defensives, tevfik, rotation, financials, overweight, valuations	investment
Topic 7	mott, cet, nim, wiles, unquestionably, sproules, nab, anz, cba, banks	banking
Topic 8	eu, theresa, brexiters, brexit, tory, boris, brussels, referendum, commons, chancellor	politics
Topic 9	monetary, ecb, kuroda, boj, qe, draghi, central, easing, bond, quantitative	economy
Topic 10	republican, republicans, democrats, democratic, trump, clinton, biden, congress, presidential, voters	politics

4.2.3. BERTopic Results

Akin to Top2Vec, the BERTopic model determines the number of topics automatically. Again, the finance researcher inspected, interpreted, and annotated a sample of the generated topics. Table 8 shows a sample of 10 topics. Figure 10 shows the c-TF-IDF scores of the keywords of a sample of topics, describing the significance of each keyword in the generated topics, i.e., how representative each word to the topic is. According to the finance researcher, the topics generated by BERTopic were better in terms of interpretability, so it was easier to annotate them with higher confidence. Figure 11 presents an inter-topic distance map of the generated topics similar to Figure 8, indicating the statistical proximity of topics.

It is worth mentioning that while BERTopic does not require a specification of how many topics to be generated, it does offer a hierarchical reduction mechanism to merge topics based on topic similarities. Figure 12 shows how some topics relate to each other, which can be used by the finance researcher to determine if they would like to merge certain topics as per their research needs.

Table 8. Topics generated by BERTopic.

No.	Top-10 Keywords	Annotation
Topic 1	wine, treasury, penfolds, wines, clarke, estates, china, brands, blass, wolf	wine
Topic 2	myer, myers, lew, umbers, premier, hounsell, lews, brookes, store, sales	retail
Topic 3	solar, energy, renewable, power, wind, grid, renewables, electricity, rooftop, projects	energy
Topic 4	afterpay, afterpays, later, merchants, pay, square, buy, eisen, molnar, credit	payment
Topic 5	china, chinese, beijing, trade, hong, xi, us, trump, kong, chinas	politics
Topic 6	climate, carbon, emissions, change, zero, warming, risks, transition, risk, climaterelated	climate
Topic 7	fed, inflation, feds, yellen, central, rates, monetary, powell, rate, us	economy
Topic 8	anz, elliott, anzs, bank, banks, banking, shayne, institutional, loans, loan	banking
Topic 9	wesfarmers, bunnings, coles, scott, goyder, homebase, kmart, gillam, stores, conglomerate	retail
Topic 10	rio, ore, iron, tonnes, rios, pilbara, mine, production, tonne, jacques	mining

4.2.4. Evaluation of Topic Models

There are no one-size-fits-all metrics for evaluating topic models. According to [16], when the TM output is used by human users, coherence scores are the most appropriate metric. There are several various coherence scores, including c_v and u_mass. The c_v score is one of the most popular coherence metrics, which builds content vectors for the words based on word co-occurrences. Then it computes the score using cosine similarity and normalized pointwise mutual information (NPMI). The u_mass score calculates how often two words appear together in the corpus, and the topic's overall coherence is determined

by averaging the pairwise coherence scores of the top N words that describe the topic. In this study, we have used both c_v and u_mass scores to evaluate the quality of topics generated by each model for comparison and validation purposes. The coherence score calculation is implemented using the Gensim Python library.

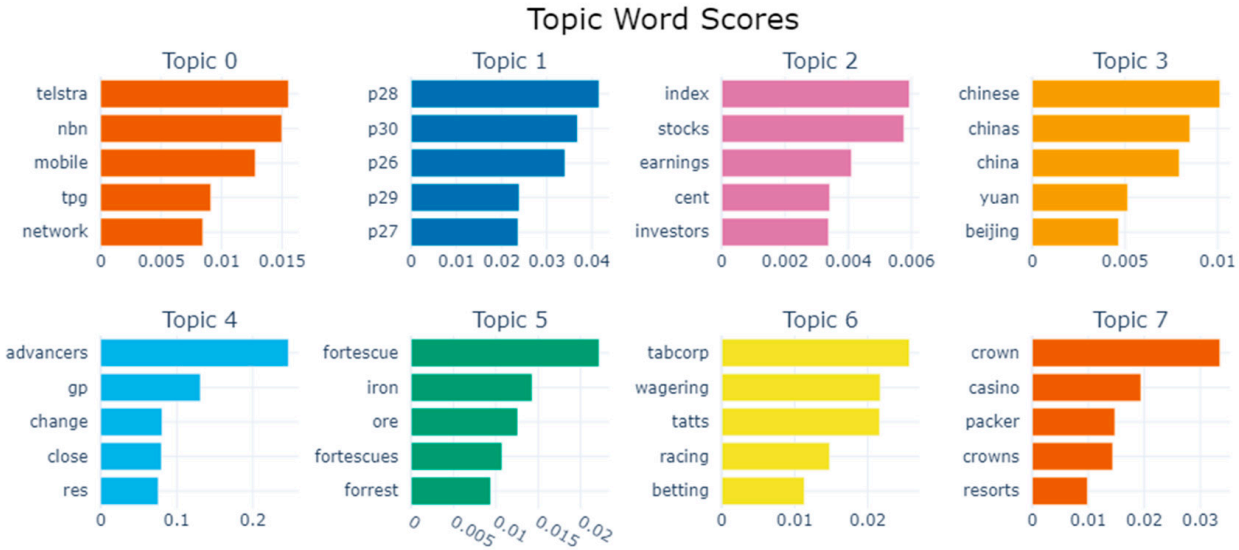

Figure 10. Topic words by the BERTopic model.

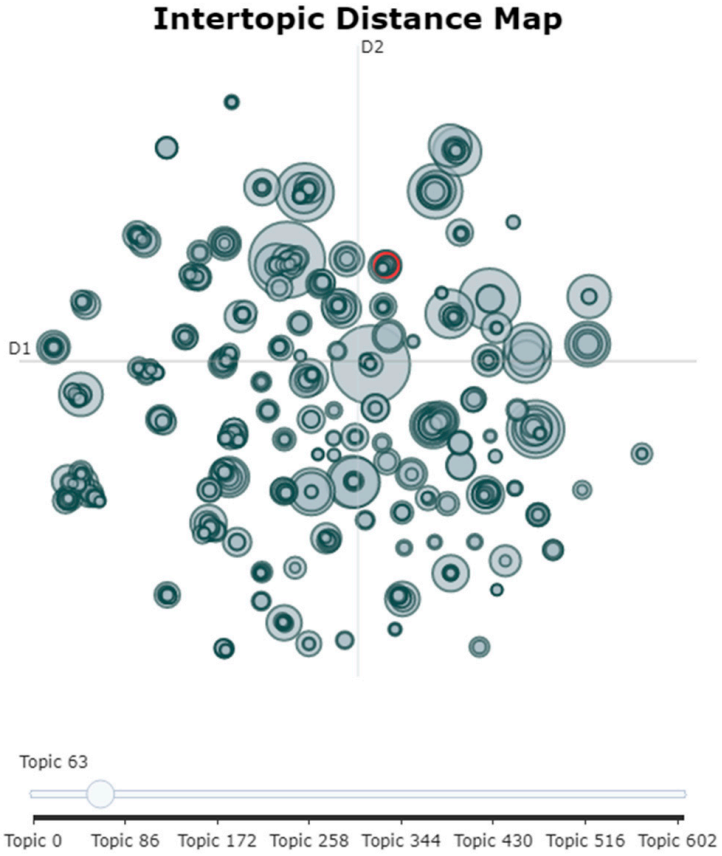

Figure 11. Visual presentation of BERTopic-generated topics.

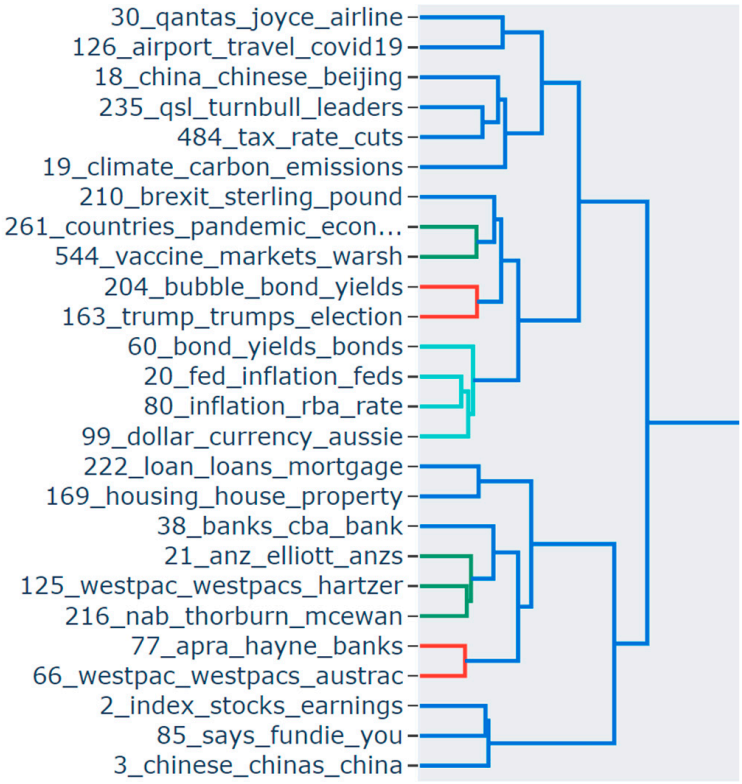

Figure 12. Part of hierarchical reduction by the BERTopic model.

Generally, the higher the coherence score is, the better the topics are. Some have argued that coherence is not a perfect measure but can be complemented by human inspection of the results for a judgment of the interpretability to verify the coherence score, which we have asked the finance expert to do in this experiment.

In addition, when dealing with many documents, the efficiency (computation time) is of great relevance. Thus, we have used the training time to evaluate the efficiency of each model.

In summary, we have used c_v and u_mass coherence scores (a quantitative method), interpretability (a subjective method), and training time to evaluate and compare the effectiveness and efficiency of each model. The detailed comparison is shown in Table 9 and Figure 13.

Table 9. Comparison between selected topic modeling models.

Topic Model	No. of Topics	Training Time (s)	Coherence (c_v)	Coherence (u_mass)
LDA 20-1	20	69.44	0.575	−16.939
LDA 50-1	50	331.99	0.582	−16.942
LDA 100-1	100	472.75	0.590	−17.197
LDA 500-1	500	777.06	0.591	−17.246
LDA 20-10	20	237.78	0.586	−17.230
LDA 50-10	50	1363.26	0.586	−17.140
LDA 100-10	100	2052.66	0.586	−17.217
LDA 500-10	500	4244.94	0.591	−17.222
Top2Vec	444	3321.50	0.545	−6.957
BERTopic	608	2632.65	0.823	−1.156

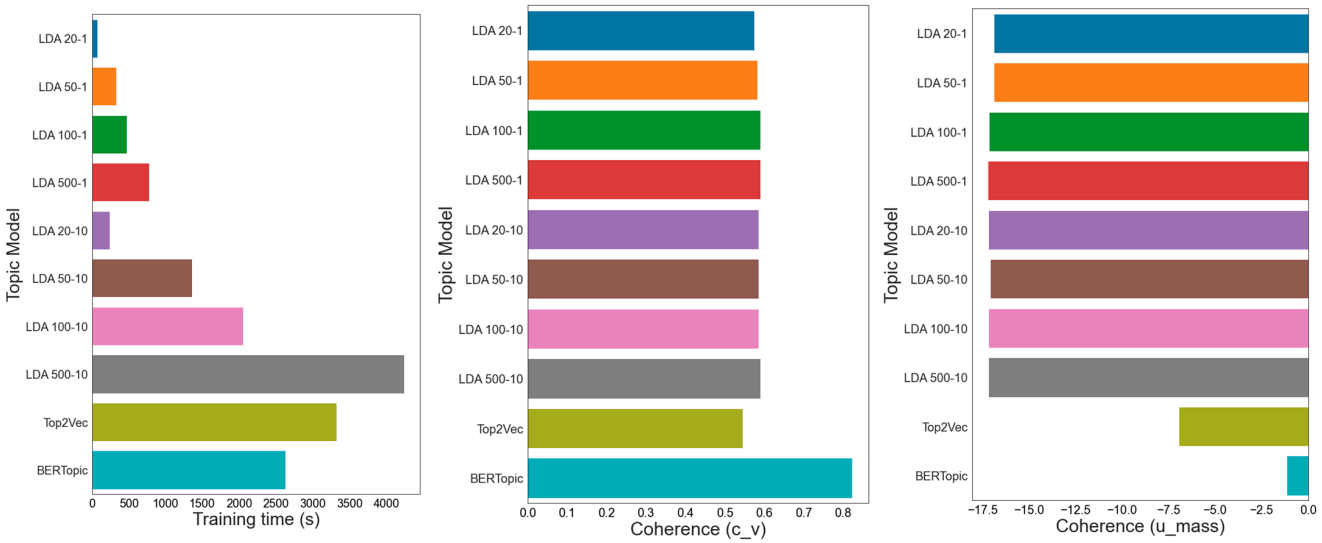

Figure 13. Comparison of various topic models used in the experiment by (**a**) training time, (**b**) c_v coherence score, and (**c**) u_mass coherence score.

LDA with a lower number of topics experienced speedier training. However, when generating a large number of topics (400–600), BERTopic was the fastest among the three models. Additionally, the c_v and u_mass coherence scores have shown that BERTopic-generated topics have higher quality than the other models. We have asked a financial researcher to inspect the generated topics of all the models and verify the results. They concluded that the coherence score makes sense, as the BERTopic model has generated topics with better interpretability. More discussion on the results will be available in Section 4.3.

4.2.5. News Impact Analysis Results

The purpose of the news impact analysis process in this study is to validate the NIA framework proposed in Section 3.1, which in turn answers RQ 2 raised in Section 1.2. It is worth noting that the news impact analysis results do not serve the purpose of evaluating topic models; instead, it demonstrates the functionality, feasibility, and usability of the NIA framework in leveraging state-of-the-art topic models for news impact analysis on financial markets.

Based on the annotation experience and the coherence scores, the finance researcher decided to use the topics generated by BERTopic. Out of the 38,240 news articles, the Topic Modeling Service selected 88 bank-related news items that belong to Topic 66, which were then fed into the RoBERTa-based Sentiment Analysis Service to generate sentiment scores, and 37 news articles with negative sentiment scores were finally selected. The Entity Extraction Service removed duplicate dates of the selected news, generating 32 unique days (events), which were then linked with the Big Four Australian banks to construct the news-entity pairs. The news selection process via topic modeling and sentiment analysis and the generation of news-entity pairs is described in Figure 14.

The financial indicator used to measure impact is mean cumulative abnormal returns (MCARs) across all stocks in question, meaning the average abnormal returns of these stocks during the analysis period defined in Table 4 (i.e., 20 days before and after the news event dates, compared with the benchmark (i.e., expected return based on the market index) in the preceding comparison period (as defined in Table 4, i.e., 100 days to 21 days prior to the news event dates). Note that the news dates are aligned and labeled as Day 0. A positive MCAR suggests a positive impact, whereas a negative MCAR suggests a negative impact. This impact analysis method is also known as the event study methodology in financial research [37].

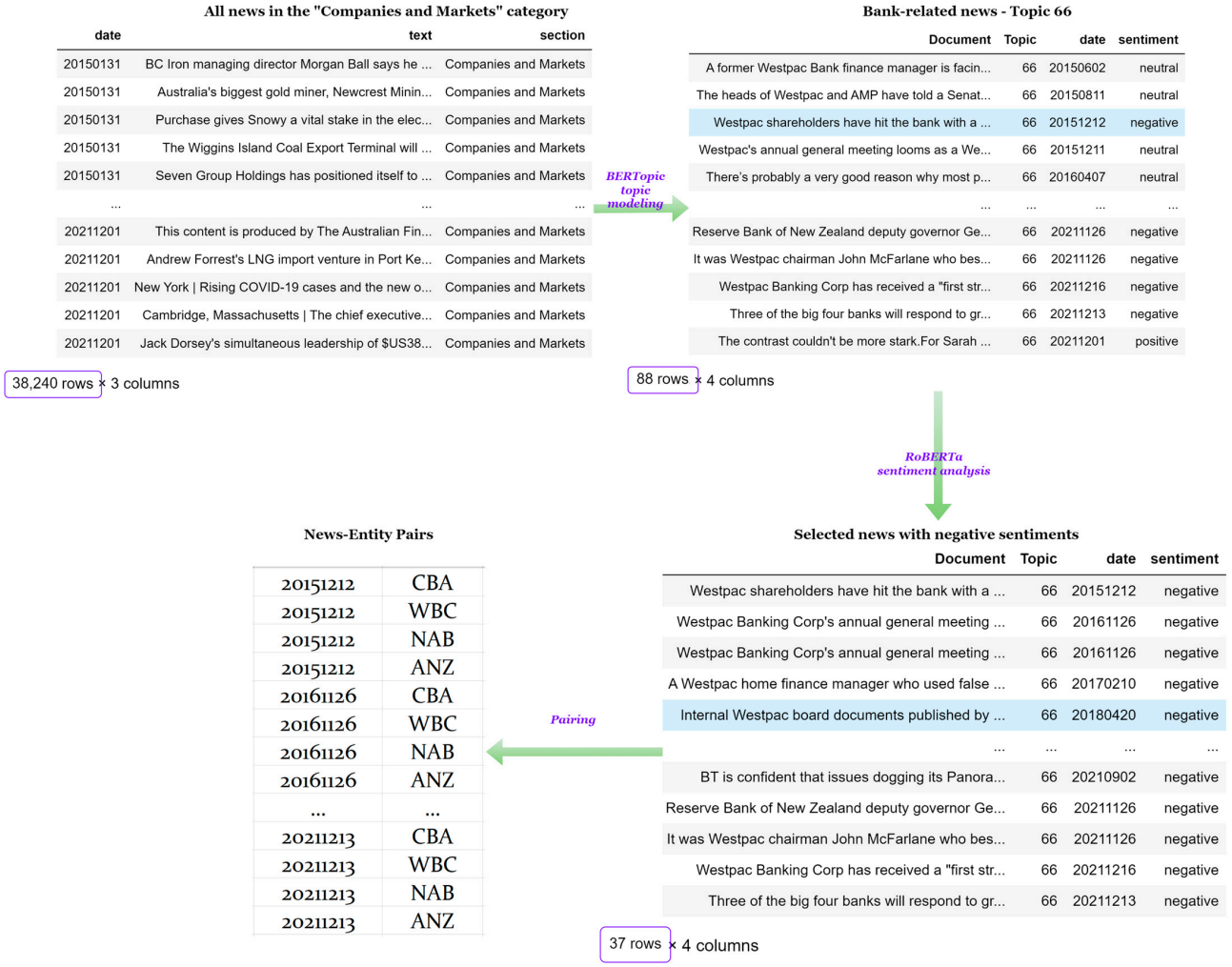

Figure 14. The news selection process based on the "banking" topic and negative sentiments.

Table 10 examines the MCARs over various window periods, showing the variation of news impact over these different periods. Figure 15 describes the MCARs discovered in this study during the analysis period. These results have been interpreted by the finance researcher based on their research assumptions and domain knowledge. One possible explanation of the observed result is that negative news related to the banking sector had negatively impacted the big four banks in Australia, as a significant drop in the mean cumulated abnormal returns was observed. However, the significant negative impact happened from 11 to 20 days before the news (MCAR = −1.210%), which may suggest proactive responses from the market due to insider knowledge before the news release. The market in general recovered after 10 days of the news date.

Table 10. MCARs over various periods.

Period (Days)	MCAR
−20 to +20	+0.165%
−10 to +10	+0.340%
−5 to +5	+0.194%
−1 to +1	+0.048%
0	+0.036%
−20 to −11	−1.210%
−10 to 0	+0.326%
0 to +10	+0.050%
+11 to +20	+1.034%

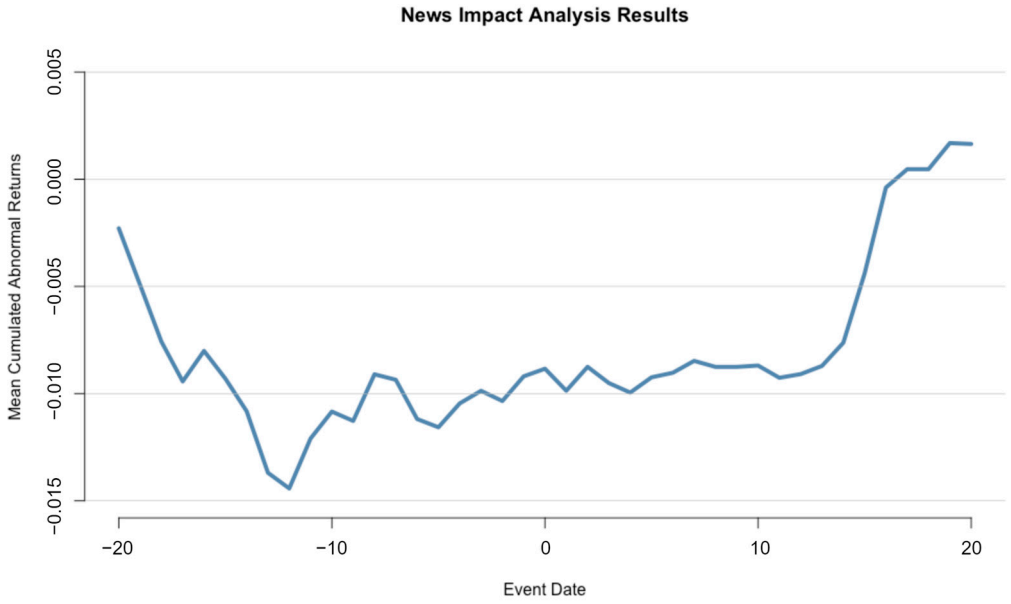

Figure 15. Mean abnormal returns identified by the news impact analysis.

4.3. Discussion

4.3.1. Comparison between LDA, Top2Vec and BERTopic

The key findings of the TM comparative results in this experimental context are as follows:

Coherence and interpretability: Out of all topic models, BERTopic by far performed best in terms of coherence, with the highest coherence scores (c_v = 0.823, u_mass = −1.156), which was in line with what the finance researcher experienced when inspecting and annotating the topics as mentioned in Section 4.2. While the c_v coherence scores of Top2Vec (0.545) were only slightly lower than those of LDA (0.586 on average), the u_mass score of Top2Vec (−6.957) was much higher than the LDA ones (−17.142 on average), regardless of the number of topics generated and the maximum iterations. The parameters of LDA (the number of topics generated and the maximum iterations) did not affect the coherence to a large degree, as the coherence scores were stable across all parameter settings.

Computation efficiency: BERTopic was faster than Top2Vec in this scenario. In comparison, the computation time of LDA relates to the number of topics generated. For a similar number of topics (500), LDA was slower than BERTopic and Top2Vec. However, LDA has the flexibility to be faster by setting a smaller number of topics. The more topics LDA generates, the more time and memory it consumes. The trade-off between coherence and resource should be considered.

Data preparation: LDA requires relatively complex data preprocessing, including removing punctuation, useless symbols, stop words, and text normalization. By contrast, one of the benefits of using Top2Vec and BERTopic is that they require minimal data preprocessing, as the underlying model needs the original structure of the text to understand the context. However, for small data samples, Top2Vec and BERTopic may generate a number of topics with some stop words. In our news impact analysis scenario, news documents tend to be long enough to mitigate this problem.

Parameter finetuning: Both Top2Vec and BERTopic utilize HDBSCAN for clustering, which does not support the specification of topic numbers to be generated by default. That said, BERTopic allows further customization by exposing the parameter settings of underlying components, including UMAP and HDBSCAN. However, LDA allows the users to flexibly determine some parameters, including the number of topics. The parameters can be finetuned using various optimization techniques such as traditional manual selection, grid search, random search, and evolutionary algorithms such as the genetic algorithm (GA).

With the above factors combined, BERTopic could be the go-to solution for analyzing long news documents' impact on financial markets such as this study. LDA can be a good alternative option to get results faster by specifying a smaller number of topics, if ready to sacrifice coherence. It is also worth noting that among the very few existing comparative studies on various topic models discussed in Section 2, ref. [31] has compared various models and concluded that BERTopic provides the highest potential with desirable results for short texts in social network studies. This comparative study focused on short texts (Tweets) only rather than long finance-related news. Thus, it is not possible to directly compare our results with theirs. However, our conclusion is overall consistent with theirs in the context of long financial news.

4.3.2. Validating the NIA Framework

As discussed in Section 4.2.5, the finance researcher who collaborated with us could complete the news impact analysis process using various services of the NIA framework, which validates the functionality, feasibility, and usability of the framework. Specifically, the finance researcher first invoked the News Import Service to obtain the original news dataset, executed the Topic Modeling Service and the Sentiment Analysis Service in turn to select news items of interest, followed by the execution of the Entity Extraction Service to generate a list of entity-news pairs (i.e., news dates and the codes of related banking stocks in this study), and the invocation of the Market Data Import Service to acquire the financial market data for related financial entities. All required data for news impact analysis was then integrated using the Data Integration Service, and finally, the finance researcher generated the analysis results using the Impact Analysis Service. Note that the Topic Modeling Service and the Sentiment Analysis Service are for news selection. Figure 14 in Section 4.2.5 has demonstrated how these services have transformed the news dataset as part of the news selection process.

It is also worth noting that without the comparison of various topic models, this entire NIA process could be fully automated. Without the NIA framework, the finance expert would have to perform most of these tasks manually as described in Figure 1. In the defined scenario, the finance expert was able to invoke the required services and complete the news impact analysis process automatically. Therefore, we can conclude that the NIA framework has facilitated a user-friendly and seamless news impact analysis process for finance researchers, with minimal IT or NLP knowledge required.

5. Conclusions and Future Work

This paper demonstrates a comparative study using three mainstream topic modeling techniques, including LDA, the most popular conventional model, and two emerging advanced neural models, namely Top2Vec and BERTopic, in the context of news impact analysis on the financial markets. We aim to discover how these topic models perform on long news articles, and how topic modeling can be integrated into a defined finance-related news impact analysis scenario. The experiment results have shown that BERTopic is overall the best-performing model, with minimal data preprocessing, the highest coherence score, and reasonable computing time. This has answered the first research question, "How do state-of-the-art TM techniques compare in the financial news context, especially with large-scale, long news corpus".

To answer the second research question, "How can topic modeling be incorporated into the financial news impact analysis framework? What is the extended process", we have advocated a news impact analysis (NIA) framework that leverages the state-of-the-art topic models wrapped into a service, to facilitate efficient news selection based on topics. The selected news can be further filtered by sentiment using a Sentiment Analysis Service, and then the news impact analysis is performed based on user-defined parameters. The experiment has validated the functionality, feasibility, and usability of the framework, which enables an automated and seamless news impact analysis process for finance domain users.

This study has a few limitations, which lead to potential future research directions. First, in the proposed NIA framework, the discovery of relevant companies or stocks and the annotation of the generated topics is semi-automated. Topics related to a particular keyword defined by the finance user are automatically used (e.g., banking), and the related companies are predefined by the finance user manually as part of the preset parameters, to generate the news-entity pairs. This can be improved by training news data labeled with relevant industrial sectors, combined with a knowledge base of mapping company codes and industrial sectors. Thus, finance experts will have a more flexible option not to specify which companies to include before the study. As part of future work, this additional function can be wrapped and added as a new service to the current version NIA framework.

Secondly, in this paper, three models (LDA, Top2Vec, and BERTopic) have been compared regarding coherence, interpretability, and computation time. Some more topics models can be compared as part of future research. In addition, as mentioned in Section 4.2.4, there is no consensus on which measure should be employed when evaluating topic models. For different applications, coherence and interpretability may not be the most relevant measure for evaluation. Other measures such as topic diversity can be of higher priority in specific applications. More research can be performed on studying and analyzing the strengths and weaknesses of various evaluation metrics.

Last but not least, the user interface of the implementation of the NIA framework can be enhanced by integrating large language models (LLMs) such as ChatGPT [38], to make the system more user-friendly and intuitive. From a different perspective, LLMs can also be used as an alternative to existing topic models, more research can be performed to compare its performance with other existing topic models.

Author Contributions: Conceptualization, W.C. and F.R.; methodology, W.C. and F.R.; software, W.C., F.R., W.L. and I.A.-Q.; validation, W.C., W.L. and I.A.-Q.; formal analysis, W.C.; investigation, W.C.; resources, W.C. and F.R.; data curation, W.C. and F.R.; writing—original draft preparation, W.C.; writing—review and editing, W.C., F.R., W.L. and I.A.-Q.; visualization, W.C.; supervision, W.C. and F.R.; project administration, W.C. and F.R.; funding acquisition, W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Fujian Province, China (Grant No. 2022J05291) and Xiamen Scientific Research Funding for Overseas Chinese Scholars.

Data Availability Statement: Restrictions apply to the availability of the news data. Daily news data is available at https://www.afr.com/, accessed on 1 May 2023, but the historical archive of news data is only available from the authors with the permission of Nine Publishing. The financial market data are sourced from Yahoo Finance and are available from the authors with the permission of Yahoo Finance.

Acknowledgments: We wish to thank Nine Publishing for giving the UNSW team access to the AFR News data used in this article.

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

References

1. Anbaee Farimani, S.; Vafaei Jahan, M.; Milani Fard, A.; Tabbakh, S.R.K. Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. *Knowl.-Based Syst.* **2022**, *247*, 108742. [CrossRef]
2. Chen, W.; El Majzoub, A.; Al-Qudah, I.; Rabhi, F.A. A CEP-driven framework for real-time news impact prediction on financial markets. *Serv. Oriented Comput. Appl.* **2023**, *17*, 129–144. [CrossRef]
3. Bonifazi, G.; Cauteruccio, F.; Corradini, E.; Marchetti, M.; Sciarretta, L.; Ursino, D.; Virgili, L. A Space-Time Framework for Sentiment Scope Analysis in Social Media. *Big Data Cogn. Comput.* **2022**, *6*, 130. [CrossRef]
4. TajMazinani, M.; Hassani, H.; Raei, R. A comprehensive review of stock price prediction using text mining. *Adv. Decis. Sci.* **2022**, *26*, 116–152.
5. Lauriola, I.; Lavelli, A.; Aiolli, F. An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing* **2022**, *470*, 443–456. [CrossRef]

6. Allen, D.E.; McAleer, M.; Singh, A.K. Daily market news sentiment and stock prices. *Appl. Econ.* **2019**, *51*, 3212–3235. [CrossRef]
7. Taj, S.; Shaikh, B.B.; Meghji, A.F. Sentiment Analysis of News Articles: A Lexicon based Approach. In Proceedings of the 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 30–31 January 2019; pp. 1–5.
8. Shahzad, F.; Yannan, D.; Kamran, H.W.; Suksatan, W.; Nik Hashim, N.A.A.; Razzaq, A. Outbreak of epidemic diseases and stock returns: An event study of emerging economy. *Econ. Res.-Ekon. Istraživanja* **2022**, *35*, 2313–2332. [CrossRef]
9. Eachempati, P.; Srivastava, P.R.; Kumar, A.; Muñoz de Prat, J.; Delen, D. Can customer sentiment impact firm value? An integrated text mining approach. *Technol. Forecast. Soc. Chang.* **2022**, *174*, 121265. [CrossRef]
10. Lin, W.-C.; Tsai, C.-F.; Chen, H. Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms. *Appl. Soft Comput.* **2022**, *130*, 109673. [CrossRef]
11. Ashtiani, M.N.; Raahemi, B. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Syst. Appl.* **2023**, *217*, 119509. [CrossRef]
12. Chen, W.; Al-Qudah, I.; Rabhi, F. A Framework for Facilitating Reproducible News Sentiment Impact Analysis. In Proceedings of the 2022 the 5th International Conference on Software Engineering and Information Management (ICSIM), Yokohama, Japan, 21–23 January 2022; pp. 125–131.
13. Bonifazi, G.; Cauteruccio, F.; Corradini, E.; Marchetti, M.; Montella, D.; Scarponi, S.; Ursino, D.; Virgili, L. Performing Wash Trading on NFTs: Is the Game Worth the Candle? *Big Data Cogn. Comput.* **2023**, *7*, 38. [CrossRef]
14. Churchill, R.; Singh, L. The Evolution of Topic Modeling. *ACM Comput. Surv.* **2022**, *54*, 215. [CrossRef]
15. Vayansky, I.; Kumar, S.A.P. A review of topic modeling methods. *Inf. Syst.* **2020**, *94*, 101582. [CrossRef]
16. Abdelrazek, A.; Eid, Y.; Gawish, E.; Medhat, W.; Hassan, A. Topic modeling algorithms and applications: A survey. *Inf. Syst.* **2023**, *112*, 102131. [CrossRef]
17. Gallagher, R.J.; Reing, K.; Kale, D.; Ver Steeg, G. Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 529–542. [CrossRef]
18. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
19. Moody, C.E. Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv* **2016**, arXiv:1605.02019.
20. Gerlach, M.; Peixoto, T.P.; Altmann, E.G. A network approach to topic models. *Sci. Adv.* **2018**, *4*, eaaq1360. [CrossRef]
21. Bhat, M.R.; Kundroo, M.A.; Tarray, T.A.; Agarwal, B. Deep LDA: A new way to topic model. *J. Inf. Optim. Sci.* **2020**, *41*, 823–834. [CrossRef]
22. Angelov, D. Top2vec: Distributed representations of topics. *arXiv* **2020**, arXiv:2008.09470.
23. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
24. Bonifazi, G.; Corradini, E.; Ursino, D.; Virgili, L. Defining user spectra to classify Ethereum users based on their behavior. *J. Big Data* **2022**, *9*, 37. [CrossRef]
25. Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A.; Pfetsch, B.; Heyer, G.; Reber, U.; Häussler, T.; et al. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Commun. Methods Meas.* **2018**, *12*, 93–118. [CrossRef]
26. Asmussen, C.B.; Møller, C. Smart literature review: A practical topic modelling approach to exploratory literature review. *J. Big Data* **2019**, *6*, 93. [CrossRef]
27. Hu, N.; Zhang, T.; Gao, B.; Bose, I. What do hotel customers complain about? Text analysis using structural topic model. *Tour. Manag.* **2019**, *72*, 417–426. [CrossRef]
28. Chen, X.; Zou, D.; Cheng, G.; Xie, H. Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. *Comput. Educ.* **2020**, *151*, 103855. [CrossRef]
29. Ghasiya, P.; Okamura, K. Investigating COVID-19 News across Four Nations: A Topic Modeling and Sentiment Analysis Approach. *IEEE Access* **2021**, *9*, 36645–36656. [CrossRef]
30. Poongodi, M.; Nguyen, T.N.; Hamdi, M.; Cengiz, K. Global cryptocurrency trend prediction using social media. *Inf. Process. Manag.* **2021**, *58*, 102708. [CrossRef]
31. Egger, R.; Yu, J. A Topic Modeling Comparison between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Front. Sociol.* **2022**, *7*, 80–92. [CrossRef]
32. Yin, H.; Song, X.; Yang, S.; Li, J. Sentiment analysis and topic modeling for COVID-19 vaccine discussions. *World Wide Web* **2022**, *25*, 1067–1083. [CrossRef]
33. Egger, R.; Yu, J. Identifying hidden semantic structures in Instagram data: A topic modelling comparison. *Tour. Rev.* **2022**, *77*, 1234–1246. [CrossRef]
34. García-Méndez, S.; de Arriba-Pérez, F.; Barros-Vila, A.; González-Castaño, F.J.; Costa-Montenegro, E. Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with Latent Dirichlet Allocation. *Appl. Intell.* **2023**. [CrossRef]
35. Alcoforado, A.; Ferraz, T.P.; Gerber, R.; Bustos, E.; Oliveira, A.S.; Veloso, B.M.; Siqueira, F.L.; Costa, A.H.R. *ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling*; Springer: Cham, Switzerland, 2022; pp. 125–136.

36. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
37. Singh, B.; Dhall, R.; Narang, S.; Rawat, S. The Outbreak of COVID-19 and Stock Market Responses: An Event Study and Panel Data Analysis for G-20 Countries. *Glob. Bus. Rev.* **2020**, 0972150920957274. [CrossRef]
38. Birhane, A.; Kasirzadeh, A.; Leslie, D.; Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **2023**, *5*, 277–280. [CrossRef]