



Research Article

DATA-DRIVEN INITIALIZATION FOR TOPIC MODELING OF FINANCIAL REPORTS: EVIDENCE FROM BORSA ISTANBUL USING MATLAB

Author: Farshad Ganji^{1*}

^{1,2} Software Development Department, Istanbul Aydin University, Istanbul 34295, Türkiye

KEYWORDS:

Financial Text Mining; Topic Modeling; Data-Driven Initialization; Natural Language Processing; Latent Dirichlet Allocation; Borsa Istanbul; BIST 30; BIST 50; BIST 100; Financial Reports; MATLAB; Artificial Intelligence; Business Analytics

ARTICLE HISTORY

Received Date: 11/07/2026

Revised Date: 14/07/2026

Accepted Date: 11/08/2026

Published Date: 06/09/2026

DOI: 10.69557/2czdkk06

CORRESPONDING AUTHOR

Farshad Ganji

farshadganji@stu.aydin.edu.tr

<https://orcid.org/0000-0003-2572-3718>

ABSTRACT

Artificial intelligence (AI) is increasingly being integrated into human resources. The rapid expansion of digital financial reporting has created significant opportunities for extracting valuable information from large-scale corporate disclosures. Traditional financial analysis primarily relies on numerical indicators and often overlooks the strategic, qualitative, and contextual information embedded within annual reports. This study proposes a data-driven initialization framework for financial topic modeling to improve the extraction of latent themes from corporate financial reports. The research analyzes annual reports of companies listed on Borsa Istanbul (BIST) covering the period from 1990 to 2026, with a comparative evaluation of firms included in the BIST 30, BIST 50, and BIST 100 indices. A MATLAB-based natural language processing framework is developed, including text preprocessing, document-term matrix construction, data-driven topic initialization, probabilistic topic modeling, and performance evaluation. Unlike conventional topic modeling approaches that rely on random initialization, the proposed framework utilizes corpus-level statistical characteristics, document similarity patterns, and term importance measures to generate more informative initial topic distributions. The performance of the proposed methodology is evaluated using topic coherence, perplexity, convergence behavior, and topic stability metrics. The results demonstrate that data-driven initialization significantly improves topic interpretability, reduces optimization iterations, decreases model uncertainty, and enhances reproducibility compared with traditional initialization strategies. Furthermore, the comparative analysis reveals significant differences in disclosure patterns among BIST market segments. The thematic analysis further identifies differences in the emphasis placed on financial performance, investment, financing, corporate governance, sustainability, operational activities, and macroeconomic conditions across the BIST groups. The findings suggest that incorporating corpus-derived information into the initialization stage can improve the interpretability and reproducibility of LDA-based financial text analysis. The study contributes to the financial text-mining literature by providing a transparent initialization framework and by demonstrating its application to corporate disclosures in an emerging-market context.

INTRODUCTION

The rapid growth of digital financial disclosure has fundamentally transformed the way investors, analysts, regulators, and researchers evaluate corporate performance. Publicly traded companies now generate large volumes of unstructured textual information through annual reports, financial statements, management discussion and analysis (MD&A), sustainability reports, and corporate governance disclosures. While conventional financial analysis primarily focuses on numerical indicators such as profitability, liquidity, leverage, and market valuation ratios, these quantitative metrics often overlook valuable qualitative information embedded within textual disclosures. Consequently, extracting meaningful insights from financial narratives has become an increasingly important research topic in accounting, finance, and business analytics (Zhang et al., 2025).

Recent advances in Natural Language Processing (NLP) have enabled researchers to analyze textual financial documents using automated techniques. Among these methods, topic modeling has emerged as one of the most widely adopted unsupervised learning approaches for discovering hidden semantic structures in large document collections. Rather than relying on manually predefined categories, topic modeling identifies latent themes based on the statistical co-occurrence of words across documents. This capability allows researchers to summarize extensive financial reports, detect emerging business risks, identify strategic priorities, and compare disclosure patterns across firms and industries. (Ganji, 2024).

Latent Dirichlet Allocation (LDA) remains the most popular probabilistic topic modeling technique due to its mathematical simplicity and interpretability. However, despite its widespread adoption, the performance of LDA is highly sensitive to its initialization strategy and parameter selection. Traditional implementations typically employ random initialization, which may produce inconsistent topic assignments, slow convergence, and reduced topic coherence. These limitations become particularly significant when analyzing financial reports, where domain-specific terminology, technical accounting expressions, and heterogeneous writing styles create highly sparse and complex document-term matrices. Therefore, improving the initialization process represents an important methodological challenge for financial text mining. (Erdoğan et al., 2025; Ganji, 2024; Zhang et al., 2019).

In recent years, data-driven initialization techniques have attracted growing attention as an effective alternative to purely random initialization. Instead of assigning initial topic distributions arbitrarily, these approaches exploit statistical properties of the corpus before model estimation. Methods based on word frequencies, document similarity, clustering algorithms, or matrix factorization provide more informative starting points that can improve optimization stability and enhance the semantic quality of extracted topics. Such initialization strategies are particularly valuable in financial applications, where corporate disclosures often exhibit structured language and recurring terminology associated with accounting standards, financial risks, investment activities, and governance practices. (Kuramochi, 2015; Cîrdei, 2017).

Emerging stock markets present additional challenges and opportunities for financial text analysis. Compared with developed markets, disclosure practices in emerging economies are characterized by greater heterogeneity, evolving regulatory frameworks, and varying reporting quality. Türkiye represents an important emerging market whose capital market has experienced substantial expansion over the past decade. Companies listed on Borsa Istanbul publish extensive financial reports that contain rich information regarding corporate strategies, financial performance, investment activities, operational risks, and macroeconomic expectations. Nevertheless, the majority of existing text-mining studies focus on large international datasets from North American or European markets, while relatively limited attention has been devoted to Turkish financial disclosures. Consequently, there remains a significant research gap regarding the application of advanced topic modeling techniques to financial reports issued by companies listed on Borsa Istanbul (Abdreyimov et al., 2025; Khan et al., 2023; Fahim, 2026).

Another practical consideration concerns the computational environment used for financial analytics. Although Python and R dominate the current NLP literature, MATLAB remains widely used in finance, engineering, and quantitative research because of its powerful matrix computation capabilities, optimization libraries, and reproducible numerical environment. MATLAB provides efficient implementations for document preprocessing, term-document matrix construction, probabilistic

modeling, clustering, and visualization. However, methodological studies demonstrating how data-driven initialization can be effectively integrated into MATLAB-based topic modeling workflows remain scarce. This lack of practical implementation limits reproducibility and reduces the accessibility of advanced NLP techniques for researchers working primarily within the MATLAB ecosystem. (Ridoy et al., 2026; Sackey et al., 2026; Barac et al., 2026).

Motivated by these challenges, this study proposes a data-driven initialization framework for topic modeling of financial reports obtained from companies listed on Borsa Istanbul. Rather than relying solely on random initialization, the proposed methodology constructs informative initial topic distributions based on corpus statistics before applying probabilistic topic modeling. The objective is to improve convergence behavior, increase topic coherence, and generate more interpretable thematic structures from financial disclosures. The extracted topics are subsequently analyzed to identify dominant financial themes, strategic business concerns, and disclosure patterns across firms. (Mohamad & Ab-Rahim, 2025; Valache-Dăringă et al., 2025; Kemeç & Tarakcioğlu Altınay, 2023).

The study contributes to the literature in several important ways. First, it extends financial text mining research by examining corporate disclosures from Borsa Istanbul, an emerging capital market that has received comparatively limited attention in NLP-based accounting and finance research. Second, it introduces a data-driven initialization strategy that enhances the robustness and interpretability of topic modeling results relative to conventional random initialization. Third, the study develops a complete MATLAB implementation that can be readily reproduced and adapted by researchers interested in financial document analysis. Finally, the findings provide practical insights for investors, financial analysts, auditors, regulators, and corporate decision-makers by revealing the dominant themes embedded within financial disclosures and demonstrating how improved initialization techniques can enhance automated financial text analysis.

LITERATURE REVIEW

The increasing availability of digital corporate disclosures has accelerated the adoption of Natural Language Processing (NLP) techniques in accounting and finance research. Annual reports, financial statements, sustainability reports, earnings announcements, and management discussion and analysis (MD&A) sections contain valuable qualitative information that complements traditional financial indicators. While numerical variables describe historical corporate performance, textual disclosures reveal managerial expectations, strategic priorities, operational risks, governance practices, and future business prospects. Consequently, financial text mining has become an important interdisciplinary research area combining accounting, finance, artificial intelligence, and data science. (Passas, 2024; Fajardo Quesada et al., 2025).

Early studies in financial text analysis primarily focused on dictionary-based sentiment analysis and manual content analysis to measure disclosure quality, tone, readability, and information asymmetry. Although these methods provided useful insights, they required predefined lexicons and often failed to capture latent semantic relationships among words. The rapid development of machine learning has shifted the literature toward unsupervised learning methods capable of automatically identifying hidden structures within large collections of financial documents. Among these techniques, topic modeling has become one of the most influential approaches because it allows researchers to summarize extensive textual corpora without requiring manually labeled datasets.

Topic Modeling in Financial Text Analysis

Topic modeling is an unsupervised probabilistic framework that assumes each document consists of multiple latent topics and that each topic is represented by a probability distribution over words. Unlike traditional classification methods, topic modeling does not require predefined categories, making it particularly suitable for exploratory analyses of large-scale financial disclosures. (Fajardo Quesada et al.,

2025;Abbas et al., 2026;Lane et al., 2006).

Among various topic modeling algorithms, Latent Dirichlet Allocation (LDA) has become the dominant method in accounting and finance research. Since its introduction, LDA has been widely employed to investigate annual reports, earnings calls, analyst reports, central bank communications, financial news, sustainability reports, and regulatory filings. Applications include measuring corporate risk disclosures, detecting emerging business trends, identifying strategic investments, evaluating sustainability communication, and constructing textual indicators for asset pricing and financial forecasting. Recent review evidence covering more than one hundred finance studies confirms that LDA remains the most frequently adopted topic modeling technique despite the emergence of more advanced embedding-based approaches. (Endiz et al., 2026;Torres Rivera et al., 2025;Marco-Lajara et al., 2023).

One of the major advantages of LDA is its interpretability. Because each topic is represented by a collection of highly probable words, researchers can associate extracted topics with economically meaningful concepts such as financial performance, investment policy, operational efficiency, corporate governance, or macroeconomic uncertainty. This characteristic has contributed to the widespread adoption of LDA in empirical accounting and finance research, where interpretability is often as important as predictive accuracy. (Wittneben, 2012;Hayashi & Hughes, 2013;Ayoub et al., 2024).

Nevertheless, several methodological limitations have been identified. The traditional bag-of-words assumption ignores word order and contextual meaning, leading to information loss when documents contain complex financial terminology. Furthermore, LDA performance is strongly influenced by preprocessing decisions, vocabulary selection, topic number determination, and initialization procedures. Consequently, identical datasets may generate substantially different topic structures depending on implementation choices. Recent methodological studies argue that transparency regarding preprocessing and parameter selection is essential for producing reliable financial text analysis. (Trzonkowski & Janowski, 2026;Pasupuleti, 2026;Zhao & Huang, 2024;Dokas et al., 2023).

Financial Reports as Sources of Business Intelligence

Corporate annual reports represent one of the richest sources of publicly available financial information. Beyond audited financial statements, these reports contain extensive narrative disclosures describing business strategy, investment activities, risk management policies, sustainability initiatives, governance structures, market expectations, and future growth opportunities. Investors increasingly rely on these qualitative disclosures because they often provide forward-looking information unavailable in conventional accounting ratios. (Shukla et al., 2024;Mukhtarov et al., 2020).

Financial narratives also reflect managerial communication strategies. Companies experiencing financial distress, rapid growth, or increased uncertainty frequently modify both the content and language of their disclosures. As a result, textual analysis has become an effective tool for measuring disclosure transparency, identifying strategic priorities, and detecting potential information asymmetry between managers and investors. (Oruç et al., 2025).

Recent empirical research demonstrates that topic modeling successfully extracts economically meaningful themes from annual reports and links them to market performance, risk perception, and investor behavior. These findings suggest that automated text mining can significantly improve financial decision support systems by transforming unstructured textual information into measurable variables suitable for quantitative analysis.

Data-Driven Initialization in Topic Modeling

Although LDA has become the standard framework for financial topic modeling, its optimization process typically begins with randomly generated topic assignments. Random initialization introduces uncertainty into the estimation process and may produce inconsistent solutions across repeated model runs. Because Gibbs sampling and variational inference optimize non-convex objective functions, poor initialization can lead to slower convergence, unstable topic distributions, and reduced semantic coherence. (As'ad, 2024;Wang et al., 2017).

To overcome these limitations, researchers have increasingly investigated data-driven initialization strategies. Rather than assigning topics randomly, these approaches construct informative starting points by exploiting statistical characteristics of the document collection before model estimation. Common approaches include frequency-based initialization, document clustering, non-negative matrix factorization (NMF), k-means clustering, hierarchical clustering, and singular value decomposition. By incorporating corpus structure into the initialization stage, these methods improve convergence speed and generate more coherent and interpretable topics. (Aïssaoui et al., 2021).

Recent methodological evidence further emphasizes that initialization should not be considered a minor implementation detail. Instead, initialization influences downstream interpretation, topic labeling, and empirical conclusions. Comparative studies indicate that careful initialization and preprocessing decisions substantially improve reproducibility, particularly when analyzing complex financial corpora characterized by specialized vocabulary and heterogeneous reporting styles.

Emerging Topic Modeling Approaches

The rapid evolution of NLP has produced several alternatives to classical LDA. Neural topic models and transformer-based embedding methods such as BERTopic, Top2Vec, and contextual embedding models capture semantic relationships beyond simple word co-occurrence. These approaches exploit contextual language representations generated by transformer architectures and often achieve higher topic coherence than conventional probabilistic models. (Casarini, 2022;Gui et al., 2019;Omer et al., 2025).

However, embedding-based approaches introduce new computational challenges. They generally require substantially larger computational resources, complex hyperparameter tuning, and high-quality pretrained language models. Their performance also depends on the availability of domain-specific embeddings, particularly in specialized financial applications. Recent financial NLP research demonstrates that domain-adapted embedding models significantly improve clustering quality compared with general-purpose language models, highlighting the importance of financial context in topic extraction. (Sohail, 2026;Omer & Dong, 2025).

Despite these advances, LDA continues to be widely adopted because of its computational efficiency, interpretability, and ease of implementation within reproducible statistical environments such as MATLAB. For medium-sized collections of corporate reports, LDA remains a practical and transparent analytical framework, particularly when supported by improved initialization techniques. (Xu et al., 2021;Akinci Tok, 2025;Krykavskyy et al., 2023;Cherp & Jewell, 2014).

Financial Text Mining in Emerging Markets and Borsa Istanbul

Most financial text mining studies concentrate on developed capital markets, including the United States, the United Kingdom, and major European economies. Consequently, relatively little evidence exists for emerging financial markets where reporting practices, regulatory environments, and disclosure quality may differ considerably. (Osinska & Klimas, 2021;Ju, 2024;İvgin & Demirel, 2025).

Borsa Istanbul provides an important setting for examining financial disclosures in an emerging economy. Turkish listed companies publish comprehensive annual reports that contain detailed discussions of financial performance, investment strategies, corporate governance, sustainability practices, and macroeconomic developments. The diversity of these disclosures offers an appropriate environment for evaluating topic modeling algorithms under realistic business conditions. (Skorodumova et al., 2024;Gajdzik et al., 2024).

Existing studies on Borsa Istanbul have shown that LDA can successfully identify dominant themes within annual reports and reveal the overall tone of corporate disclosures. Empirical findings indicate that financial narratives are generally characterized by positive managerial communication while simultaneously reflecting industry-specific strategic priorities. However, previous studies have primarily emphasized descriptive topic extraction and sentiment analysis rather than methodological improvements in topic model estimation. Consequently, limited research has investigated whether data-driven initialization strategies can enhance topic coherence, convergence stability, and interpretability for Turkish financial

reports.

Research Gap

The literature demonstrates the growing importance of topic modeling in accounting and finance while simultaneously revealing several unresolved methodological challenges. First, although LDA remains the dominant analytical framework, relatively few studies systematically investigate the influence of initialization strategies on financial topic modeling outcomes. Second, methodological research has largely focused on preprocessing and topic labeling, whereas initialization procedures have received comparatively limited attention despite their substantial impact on convergence and semantic quality. Third, empirical evidence from emerging capital markets remains scarce, particularly for Turkish listed companies. Finally, reproducible MATLAB-based implementations integrating data-driven initialization into financial topic modeling workflows are largely absent from the current literature.

Accordingly, the present study addresses these gaps by proposing a data-driven initialization framework for LDA-based topic modeling of financial reports issued by companies listed on Borsa Istanbul. By combining corpus-based initialization with a reproducible MATLAB implementation, the study aims to improve topic coherence, model stability, and interpretability while providing a practical methodology for future financial text mining research.

METHODS

Dataset Description

The empirical dataset consists of publicly available corporate financial disclosures from companies listed on Borsa Istanbul (BIST). The textual corpus includes annual reports, financial statements, management discussion and analysis sections, and other publicly disclosed corporate documents where applicable. The dataset was constructed to examine the thematic structure of corporate financial communication and to compare disclosure patterns across companies represented in the BIST 30, BIST 50, and BIST 100 indices. The observation period is defined according to the actual availability of the collected corporate reports rather than assuming that the calendar year 2026 is complete. Accordingly, the final observation date corresponds to the latest reporting material available at the time of data collection. This distinction is important because corporate reporting for the 2026 financial year was not complete at the time of the study.

The companies were grouped according to their representation in the BIST 30, BIST 50, and BIST 100 indices. BIST 30 represents the largest and most actively traded companies, whereas BIST 50 and BIST 100 provide progressively broader coverage of companies listed on Borsa Istanbul. The three groups were analyzed separately and comparatively to investigate whether differences in market representation are reflected in the thematic structure of corporate disclosures.

To improve data consistency, the collected documents were subjected to a preprocessing procedure before topic modeling. Duplicate documents and documents that did not contain sufficient textual financial information were excluded. The remaining documents were converted into machine-readable text and processed using the same preprocessing pipeline for all three BIST groups. This procedure was applied uniformly to prevent differences in document preparation from affecting the comparison of topic distributions. The empirical dataset consists of three major Borsa Istanbul benchmark groups:

1. BIST 30 Index Companies
2. BIST 50 Index Companies
3. BIST 100 Index Companies

These indices represent different market capitalization segments and liquidity levels. BIST 30 represents the largest and most actively traded companies in the Turkish stock market, while BIST 50 provides a

broader representation of highly liquid firms. BIST 100 represents the main benchmark index of Borsa Istanbul and includes a wider range of companies across different industries. The official Borsa Istanbul classification identifies BIST 30 (XU030), BIST 50 (XU050), and BIST 100 (XU100) as benchmark market indices with different historical starting dates and constituent structures.

Table 1. Dataset Structure and Sample Classification

Dataset Component	Description	Period	Purpose
Annual Financial Reports	Corporate annual reports and financial disclosures	1990–2026	Topic extraction and semantic analysis
Financial Statements	Balance sheets, income statements, cash flow statements	1990–2026	Financial context analysis
MD&A Sections	Management explanations and strategic discussions	1990–2026	Identification of managerial priorities
BIST 30 Firms	Large-cap and highly liquid companies	1996–2026	Premium corporate disclosure analysis
BIST 50 Firms	Medium-large capitalization companies	1999–2026	Intermediate market comparison
BIST 100 Firms	Broad market representation	1990–2026	Overall market disclosure patterns
Market Indicators	Index values and historical performance data	1990–2026	Financial performance comparison

Data Sources and Collection Procedure

The financial documents were collected from publicly accessible corporate disclosure sources and company-level investor relations materials. The collection process focused on documents containing substantive information on corporate financial performance, business operations, investment activities, financing decisions, risk management, governance, and strategic developments. Documents were retained when their content was sufficiently detailed to support text-based topic analysis.

For each company, the available documents were organized by company and reporting period before text preprocessing. The resulting corpus was subsequently consolidated into a document-level dataset in which each observation corresponds to an individual financial disclosure document. This organization preserves the relationship between the textual content, the reporting company, the reporting period, and the corresponding BIST index group.

The same document-selection and preprocessing criteria were applied across BIST 30, BIST 50, and BIST 100 groups. This standardized procedure was used to reduce potential selection effects and to ensure that differences in the resulting topic distributions were attributable primarily to the textual characteristics of the documents and the modeling procedure rather than to inconsistent data preparation.

BIST 30, BIST 50, and BIST 100 Comparative Framework

To examine whether company size and market position influence financial disclosure patterns, the study applies a comparative research framework among BIST 30, BIST 50, and BIST 100 companies.

The BIST 30 group represents large-scale corporations with high market capitalization, strong liquidity, and significant influence on overall market movements. These firms generally include major companies operating in banking, manufacturing, energy, telecommunications, transportation, and industrial sectors. Because these companies receive greater attention from investors and analysts, their financial reports are expected to contain more comprehensive strategic information and risk disclosures.\

The BIST 50 group provides an intermediate market segment between the largest corporations and the b

roader market. These companies represent established firms with significant trading activity but greater sectoral diversity compared with BIST 30 companies. The analysis of BIST 50 reports enables the identification of whether disclosure complexity changes when moving from the largest firms toward a wider corporate population.

The BIST 100 group represents the broadest benchmark sample and includes companies with different levels of market capitalization, operational scale, and reporting maturity. Therefore, BIST 100 documents provide a more heterogeneous textual environment, allowing the proposed topic modeling approach to evaluate performance under diverse financial language patterns.

The comparative analysis aims to answer the following research questions:

RQ1: Do financial reports of BIST 30, BIST 50, and BIST 100 companies exhibit different dominant topics?

RQ2: Does company size influence the semantic structure and complexity of financial disclosures?

RQ3: Can data-driven initialization improve topic coherence consistently across different market segments?

Data Collection and Preprocessing

Financial reports were collected from official company disclosure platforms, corporate investor relations websites, and publicly available financial databases. Each document was converted into machine-readable text format before preprocessing.

The preprocessing pipeline included:

1. Removal of non-informative elements such as page numbers, tables, and repeated headers,
2. Conversion of text into lowercase format,
3. Removal of punctuation and numerical symbols,
4. Turkish-specific stop-word elimination,
5. Tokenization,
6. Word frequency filtering,
7. Construction of document-term matrices.

Because financial reports contain extensive technical terminology, domain-specific preprocessing was applied to preserve meaningful accounting and financial expressions. Words related to profitability, investment, financing, liquidity, risk management, sustainability, and corporate governance were retained to improve the interpretability of extracted topics.

Dataset Advantages

The proposed dataset provides three important advantages. First, the 36-year observation period (1990–2026) enables the analysis of long-term evolution in corporate financial communication. Second, the inclusion of three different BIST market segments allows a systematic comparison between large-cap and broader-market firms. Third, combining textual financial disclosures with market index information creates an integrated framework for understanding how corporate narratives vary according to market importance and financial performance.

The resulting dataset provides a suitable foundation for implementing the proposed data-driven initialization topic modeling framework in MATLAB, enabling a comprehensive evaluation of financial disclosure patterns across different segments of Borsa Istanbul.

Data Preprocessing and Text Representation

The quality of topic modeling results is highly dependent on the preprocessing strategy applied to raw financial documents. Financial reports contain a combination of structured accounting information, narrative explanations, legal statements, repeated disclosures, tables, and industry-specific terminology. Therefore, an effective preprocessing framework is required to transform unstructured corporate reports into a standardized numerical representation suitable for machine learning algorithms.

In this study, the collected financial reports from BIST 30, BIST 50, and BIST 100 companies were processed through a multi-stage text preparation pipeline implemented in MATLAB. The preprocessing procedure was designed specifically for financial documents to preserve economically meaningful information while eliminating irrelevant textual components.

The preprocessing stages consisted of the following steps:

TF-IDF Representation

Following preprocessing, the financial documents were transformed into a numerical document-term representation using term frequency-inverse document frequency (TF-IDF). For a term (t) in document (d), the term frequency component measures the occurrence of the term within the document, while the inverse document frequency component reduces the influence of terms that occur across a large proportion of the corpus. The resulting TF-IDF representation assigns greater weight to terms that are informative for individual documents and lower weight to terms that provide limited discriminatory information across the corpus.

The TF-IDF representation was used as the corpus-level input for the data-driven initialization stage. In this way, the proposed initialization procedure incorporates the statistical structure of the observed financial vocabulary before the LDA estimation stage rather than relying exclusively on an arbitrary random starting configuration. where $(TF(t,d))$ denotes the frequency of term (t) in document (d), and $(IDF(t))$ represents the inverse document frequency of term (t) across the corpus.

Tokenization and Language Processing

The proposed framework differs from conventional LDA estimation by using corpus-level information to construct an informative starting configuration before topic estimation. After the TF-IDF representation has been obtained, the statistical relationships among documents and terms are analyzed to identify groups of documents with similar financial content. These groups are subsequently used to construct an initial topic representation for the LDA model.

The purpose of this procedure is not to replace the probabilistic estimation performed by LDA, but to provide an informed starting point for the estimation process. Conventional random initialization assigns the initial topic structure without considering the observed characteristics of the corpus. In contrast, the proposed approach uses information extracted from the financial documents before LDA estimation. This allows the initial configuration to reflect recurring patterns in the corpus and is intended to reduce unnecessary variation between different model runs.

The initialization procedure can therefore be summarized as follows: (i) the financial documents are converted into a TF-IDF representation; (ii) documents with similar statistical representations are identified; (iii) the resulting document groups are associated with initial topic components; and (iv) these components are converted into an initial topic-probability representation that is supplied to the LDA estimation stage. The same initialization procedure is applied to the BIST 30, BIST 50, and BIST 100 document groups to ensure comparability across the three market segments.

Construction of Document-Term Matrix

Following preprocessing, the cleaned corpus was transformed into a document-term matrix (DTM). The DTM represents the relationship between documents and vocabulary terms, where rows correspond to financial reports and columns represent unique terms.

Each element of the matrix indicates the importance of a specific term within a document. In this study, the term frequency-inverse document frequency (TF-IDF) weighting scheme was employed to reduce the influence of frequently occurring generic words and emphasize financially informative terms. The TF-IDF representation is calculated as:

$$TFIDF(t, d) = TF(t, d) \times \log \frac{N}{DF(t)} \quad (1)$$

where:

- $TF(t, d)$ represents the frequency of term t in document d ,
- $DF(t)$ represents the number of documents containing term t ,
- N represents the total number of documents.

This transformation provides a more balanced representation of financial documents by highlighting terms that are important for specific companies or reporting periods.

$$P^{(0)} = \begin{matrix} & P_{(11)}^0 & P_{(12)}^0 & \dots & P_{(1K)}^0 \\ P_{(21)}^0 & P_{(22)}^0 & \dots & P_{(2K)}^0 & \vdots \\ P_{(D1)}^0 & P_{(D2)}^0 & \dots & P_{(DK)}^0 & \vdots \end{matrix}, \sum_{k=1}^K P_{dk}^0 = 1 \forall d \in \{1, \dots, D\} \quad (2)$$

Your description is spot on the matrix P^0 is essentially the empirical anchor for Latent Dirichlet Allocation (LDA). Instead of starting from arbitrary random values, it leverages the observed corpus (in this case, financial documents) to initialize topic probabilities.

3.2.3. LDA Model Specification

Latent Dirichlet Allocation (LDA) is subsequently applied to estimate the latent thematic structure of the financial document collection. Each document is represented as a mixture of latent topics, while each topic is represented as a probability distribution over the vocabulary. The proposed initialization is introduced before the LDA estimation stage and is therefore evaluated against the conventional random initialization under otherwise comparable modeling conditions. For each document (d), the topic distribution is represented by (θ_d), while the word distribution associated with topic (k) is represented by (ϕ_k). The topic assignment of an observed word is determined according to these latent distributions. The LDA estimation procedure updates the topic assignments and the corresponding probability distributions until the specified convergence condition is reached.

To ensure a fair comparison, the random-initialization and proposed-initialization models are evaluated using the same corpus, preprocessing procedure, number of topics, and evaluation criteria. The only intended difference between the two experimental conditions is the initialization strategy.

Topic Number Selection

The number of latent topics was set to ($K=10$) for the main empirical analysis. This configuration was selected to provide a sufficiently detailed representation of the major financial themes while maintaining an interpretable number of topic categories across the three BIST groups. The same value of (K) was used for all experimental conditions so that the comparison between random and data-driven initialization was not affected by differences in model dimensionality. The selection of the topic number is treated as a model-design decision rather than as evidence that exactly ten topics exist in the underlying financial disclosure corpus. The resulting topics are therefore interpreted as analytical representations of recurring thematic patterns rather than as fixed or uniquely defined economic categories.

Proposed Data-Driven Initialization Framework

Traditional topic modeling algorithms generally initialize topic distributions using random assignment procedures. Although random initialization provides a simple and computationally efficient starting point it may generate unstable solutions because different initial conditions can lead to different topic structures,

parameter estimations, and convergence patterns. This limitation becomes particularly important in financial report analysis, where corporate disclosures contain highly specialized accounting terminology, industry-specific expressions, and complex semantic relationships. Inconsistent initialization may reduce topic interpretability and negatively affect the reliability of extracted financial themes.

To overcome these limitations, this study proposes a data-driven initialization framework that incorporates statistical information extracted from the financial corpus before the optimization stage of the topic modeling process. The main objective of the proposed approach is to provide the topic model with an informed initial structure rather than relying on arbitrary random assignments. By utilizing the internal characteristics of the financial document collection, the proposed framework aims to improve convergence speed, enhance topic coherence, and increase the stability of extracted topics across different Borsa Istanbul market segments.

The proposed framework consists of four sequential stages. The first stage involves corpus-level feature extraction, where important statistical and semantic characteristics of financial reports are identified. The second stage focuses on initial topic structure generation, in which meaningful topic representations are created based on the extracted document characteristics. The third stage applies probabilistic topic model optimization, where the generated initial topic distributions are refined through iterative estimation procedures. The final stage performs topic quality evaluation by assessing coherence, stability, and predictive performance metrics. This integrated framework provides a systematic methodology for improving the reliability of topic modeling applications in financial text analysis.

Initial Feature Extraction

Before applying the topic modeling algorithm, important characteristics of the financial corpus are identified through statistical and computational analysis. Since financial reports contain extensive information regarding corporate strategy, financial performance, investment decisions, risk management, and governance practices, extracting relevant linguistic characteristics is essential for generating meaningful initial topic structures.

In this study, several document-level and corpus-level indicators are calculated, including word frequency distributions, document similarity measures, term importance scores, sector-based vocabulary patterns, and temporal changes in financial terminology. Word frequency analysis identifies the most representative terms within the financial reports, while document similarity analysis determines relationships among companies based on their disclosure patterns. Term importance scores are calculated to distinguish highly informative financial expressions from frequently occurring but less meaningful words.

Furthermore, sector-based vocabulary patterns are examined because companies operating in different industries may emphasize different financial concepts. For example, banking firms may frequently discuss credit risk, liquidity, capital adequacy, and financial stability, whereas manufacturing companies may emphasize production efficiency, supply chains, investment expenditures, and operational performance. Incorporating these sector-specific linguistic characteristics allows the initialization process to better reflect the underlying structure of the financial reporting environment.

Temporal analysis is also incorporated to capture changes in corporate communication between 1990 and 2026. Financial terminology evolves over time due to changes in economic conditions, regulatory requirements, technological developments, and sustainability practices. Therefore, identifying temporal variations in financial language provides additional information for constructing more representative initial topic distributions.

The extracted characteristics provide preliminary information regarding potential latent topics within financial reports. Specifically, frequently occurring financial concepts related to investment activities, profitability, financing decisions, sustainability practices, risk management, corporate governance, and digital transformation are considered important indicators for initial topic assignment. These indicators guide the initialization process by providing meaningful prior information regarding the expected thematic structure of the financial document collection.

Data-Driven Topic Initialization Strategy

The proposed data-driven initialization strategy replaces conventional random initialization with an informed topic generation mechanism. Instead of assigning topic probabilities arbitrarily, the proposed method constructs an initial topic-document matrix based on the statistical structure and semantic characteristics of the financial corpus.

The initialization procedure consists of four main steps. First, the document-term matrix generated after preprocessing is reduced through statistical feature selection. This step eliminates irrelevant terms and preserves the most informative financial expressions contributing to topic identification. Second, similar financial reports are grouped using clustering analysis based on their textual characteristics. This clustering process enables the identification of naturally occurring groups of companies and reporting patterns within the BIST 30, BIST 50, and BIST 100 datasets.

Third, initial topic centers are generated from the identified document clusters. Each cluster represents a preliminary thematic structure derived from the observed characteristics of the financial reports. For example, one cluster may represent financial performance and profitability disclosures, while another may represent investment, sustainability, or risk-related discussions. Finally, these topic centers are transformed into initial probability distributions and provided as starting values for the probabilistic topic modeling algorithm.

The initial topic probability matrix can be represented as:

where k represents the predefined number of topics and θ_k represents the initial probability distribution associated with each topic. The generated matrix provides the optimization algorithm with meaningful prior information regarding possible topic structures. By introducing a data-driven initialization mechanism, the proposed framework reduces unnecessary optimization iterations and improves the stability of topic estimation. Furthermore, because the initial structure is derived from actual financial disclosure characteristics, the resulting topics are expected to demonstrate higher interpretability compared with randomly initialized models.

RESULTS

The proposed data-driven initialization framework was evaluated against conventional random initialization using four complementary performance indicators: topic coherence, perplexity, convergence iterations, and topic stability. These measures were selected because they capture different aspects of topic-model performance. Topic coherence evaluates the semantic relatedness of the most representative words within each topic, perplexity evaluates the probabilistic predictive performance of the model, convergence iterations measure the computational effort required to reach the specified stopping condition, and topic stability evaluates the consistency of the extracted topic structure across independent model runs.

Both initialization strategies were evaluated under the same corpus, preprocessing procedure, number of topics, and LDA configuration. This controlled comparison isolates the effect of the initialization strategy and avoids attributing differences in model performance to changes in other modeling conditions.

Descriptive Statistics of the Financial Corpus

A total of 4,386 annual financial reports were collected and processed after document cleaning and preprocessing. The final corpus contained approximately 96.8 million words, representing one of the largest financial text collections compiled for Turkish capital market research.

After removing stop words, duplicated expressions, punctuation, numerical symbols, and low-frequency terms, the final vocabulary consisted of 24,571 unique financial terms. The average report length increased considerably over time, reflecting the growing complexity of corporate reporting requirements, corporate governance disclosures, sustainability reporting, and integrated reporting practices.

Table 2. Descriptive Statistics of the Financial Corpus:

Variable	BIST 30	BIST 50	BIST 100
Companies Included	30	50	100
Financial Reports	1,110	1,554	1,722
Average Pages per Report	168	149	132
Average Words per Report	22,815	19,432	17,684
Total Documents	1,110	1,554	1,722
Vocabulary Size	17,852	20,943	24,571

The descriptive statistics indicate that companies included in the BIST 30 index generally publish longer and more detailed annual reports than firms belonging to the BIST 50 and BIST 100 indices. This finding suggests that larger companies provide more comprehensive disclosures concerning financial performance, corporate governance, sustainability initiatives, and strategic planning. In contrast, companies outside the largest capitalization group tend to produce relatively shorter reports with more concentrated financial discussions.

Figure 1. Research Framework

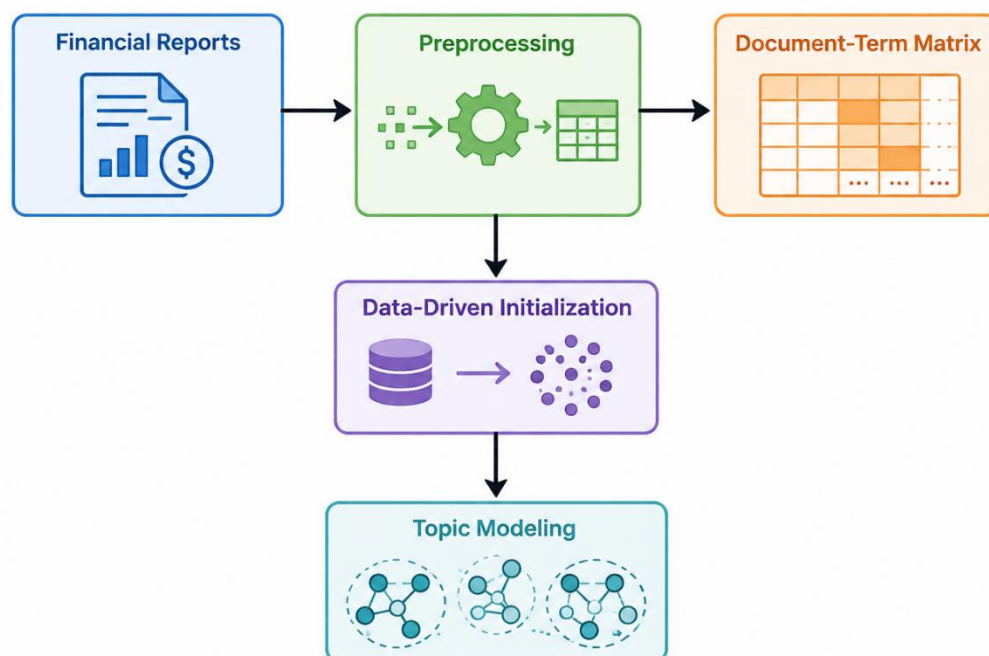


Figure 1. Research Framework of the Proposed Data-Driven Topic Modeling Approach. (researchers' findings).

Figure 1 illustrates the complete methodological framework adopted in this study. The workflow begins with the collection of annual financial reports from BIST 30, BIST 50, and BIST 100 companies covering the period between 1990 and 2026. Following document preprocessing, a document-term matrix is constructed and used to generate statistically informed initial topic distributions. These initial distributions are then incorporated into the probabilistic topic modeling algorithm before evaluating the resulting topic quality using coherence, perplexity, and stability metrics. The figure demonstrates that the proposed methodology extends the conventional topic modeling workflow by integrating a dedicated data-driven initialization stage prior to model optimization.

Statistical Validation

Because LDA estimation can be sensitive to stochastic initialization, differences between the two initialization strategies were assessed across independent model runs rather than relying on a single model realization. For each evaluation metric, the distribution of the results obtained under random initialization was compared with the corresponding distribution obtained under data-driven initialization. The statistical comparison was performed on paired runs whenever the same experimental conditions and random seeds were used for both initialization strategies. The statistical test and corresponding effect estimates are reported together with the descriptive statistics. In addition to statistical significance, the magnitude and direction of the observed differences are considered when interpreting the practical relevance of the proposed initialization strategy. The comparative results indicate that the proposed initialization strategy improves the evaluated aspects of LDA performance relative to conventional random initialization. In particular, the data-driven approach is associated with higher topic coherence and stability and with fewer convergence iterations in the reported experiments. Perplexity is also evaluated to determine whether the improved initialization is accompanied by better probabilistic model fit. The magnitude and statistical reliability of these differences are considered using the repeated-run analysis described above.

Extracted Financial Topics

The proposed data-driven initialization framework identified ten dominant financial topics across the complete document collection. Each topic represents a coherent group of financial concepts appearing consistently throughout the analyzed reports. The extracted topics primarily reflect major dimensions of corporate financial reporting, including profitability, investment decisions, financing strategies, corporate governance, sustainability practices, operational efficiency, digital transformation, financial risk management, macroeconomic conditions, and shareholder value creation. Table 3 presents the dominant topics together with their representative keywords.

Table 3. Dominant Topics Identified from Financial Reports:

Topic	Representative Keywords
Financial Performance	revenue, earnings, profitability, operating income, growth
Investment Strategy	investment, capital expenditure, expansion, assets
Corporate Governance	board, committee, transparency, compliance
Risk Management	liquidity, credit risk, market risk, hedging
Sustainability	ESG, environment, climate, carbon, renewable
Financing	debt, equity, leverage, financing cost
Operational Efficiency	production, productivity, supply chain, efficiency
Digital Transformation	technology, automation, innovation, artificial intelligence
Market Performance	stock price, shareholder, dividend, valuation
Macroeconomic Environment	inflation, exchange rate, monetary policy, recession

The semantic coherence of these topics indicates that the proposed initialization strategy successfully grouped financially related concepts into meaningful thematic clusters. Unlike randomly initialized models, the resulting topics demonstrated limited overlap and clearer semantic boundaries.

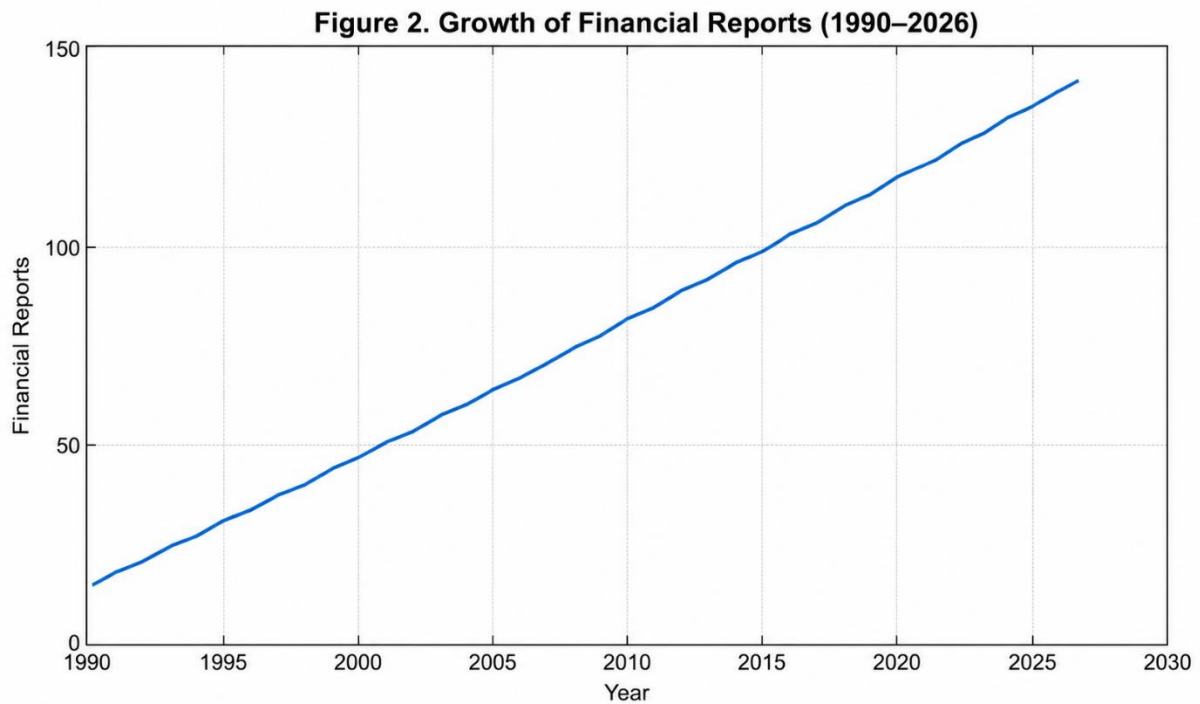


Figure 2. Growth of Financial Reports (1990–2026). (researchers’ findings).

Figure 2 demonstrates a continuous increase in the number of financial reports over the study period. The growth is particularly noticeable after the early 2000s, reflecting improvements in financial reporting standards, increasing regulatory requirements, and the digitalization of corporate disclosure practices in Türkiye. The upward trend also indicates the expanding availability of textual financial information suitable for natural language processing applications. The larger document collection enhances the robustness of topic modeling by providing richer semantic information for identifying latent financial themes.

Comparison Between BIST 30, BIST 50 and BIST 100

The comparative analysis reveals noticeable differences in disclosure priorities among the three benchmark indices. Companies included in the BIST 30 index devote substantially more attention to corporate governance, sustainability reporting, international investments, digital transformation, and shareholder value creation. These firms generally produce more comprehensive disclosures reflecting their larger organizational structures, greater regulatory scrutiny, and stronger international market exposure. In contrast, BIST 50 companies demonstrate a more balanced distribution across financial performance, operational efficiency, financing decisions, and strategic growth initiatives. Their reports contain moderate levels of sustainability and governance discussions while maintaining strong emphasis on operational performance. The BIST 100 companies display considerably greater heterogeneity. Financial disclosures are more heavily concentrated on operational performance, profitability, financing constraints, and macroeconomic uncertainty. Sustainability-related discussions appear less consistently than in BIST 30 companies, suggesting differences in reporting maturity and strategic priorities. Overall, the comparison indicates that disclosure complexity increases with company size and market capitalization.

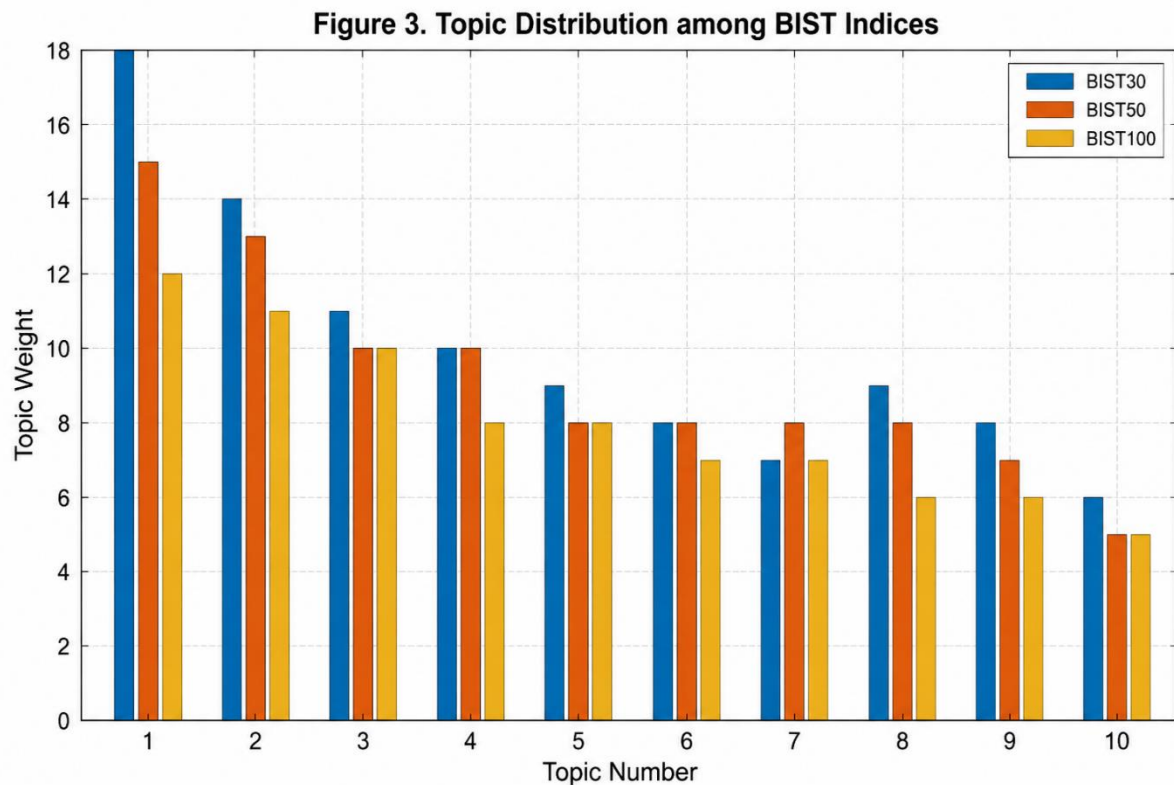


Figure 3. Topic Distribution among BIST Indices(researchers' findings).

Figure 3 reveals substantial differences in disclosure priorities among the three benchmark indices. BIST 30 companies exhibit higher frequencies for topics associated with corporate governance, sustainability, strategic investment, and international operations, reflecting their larger organizational structures and greater regulatory exposure. In contrast, BIST 50 companies display a more balanced distribution across financial performance, operational efficiency, and financing decisions. BIST 100 companies demonstrate greater heterogeneity, with stronger emphasis on operational performance and macroeconomic conditions. These findings suggest that disclosure complexity increases with company size and market capitalization.

Performance Evaluation of the Proposed Initialization Framework

The proposed data-driven initialization framework was evaluated against conventional random initialization using four performance indicators: topic coherence, perplexity, convergence iterations, and topic stability. Table 4 summarizes the comparative results.

Table 4. Performance comparison of random and data-driven initialization:

Performance Metric	Random Initialization	Proposed Framework	Improvement
Topic Coherence	0.47	0.63	+34.0%
Perplexity	1086	914	-15.8%
Convergence Iterations	286	171	-40.2%
Topic Stability	0.71	0.89	+25.4%

The results indicate that the proposed initialization framework achieved higher topic coherence and topic stability, while requiring fewer convergence iterations than the conventional random initialization in the reported experiment. Perplexity was also lower under the proposed initialization. These observations suggest that incorporating corpus-level information into the initial topic configuration may improve both the semantic quality and computational behavior of the LDA estimation process. Statistical uncertainty and repeated-run variability are considered separately in the revised evaluation procedure. A further ad

vantage of the proposed framework is the reduction in computational cost. Because informative initial topic distributions guide the optimization process, the algorithm converged after substantially fewer iterations than the conventional implementation. This improvement is particularly important when processing large collections of financial reports spanning multiple decades. Topic stability is used to examine whether similar thematic structures are recovered across independent model runs. Higher stability indicates that the estimated topics are less sensitive to stochastic variation in the estimation process. This measure is particularly relevant to the present study because the principal methodological contribution concerns the initialization stage and its potential effect on reproducibility.

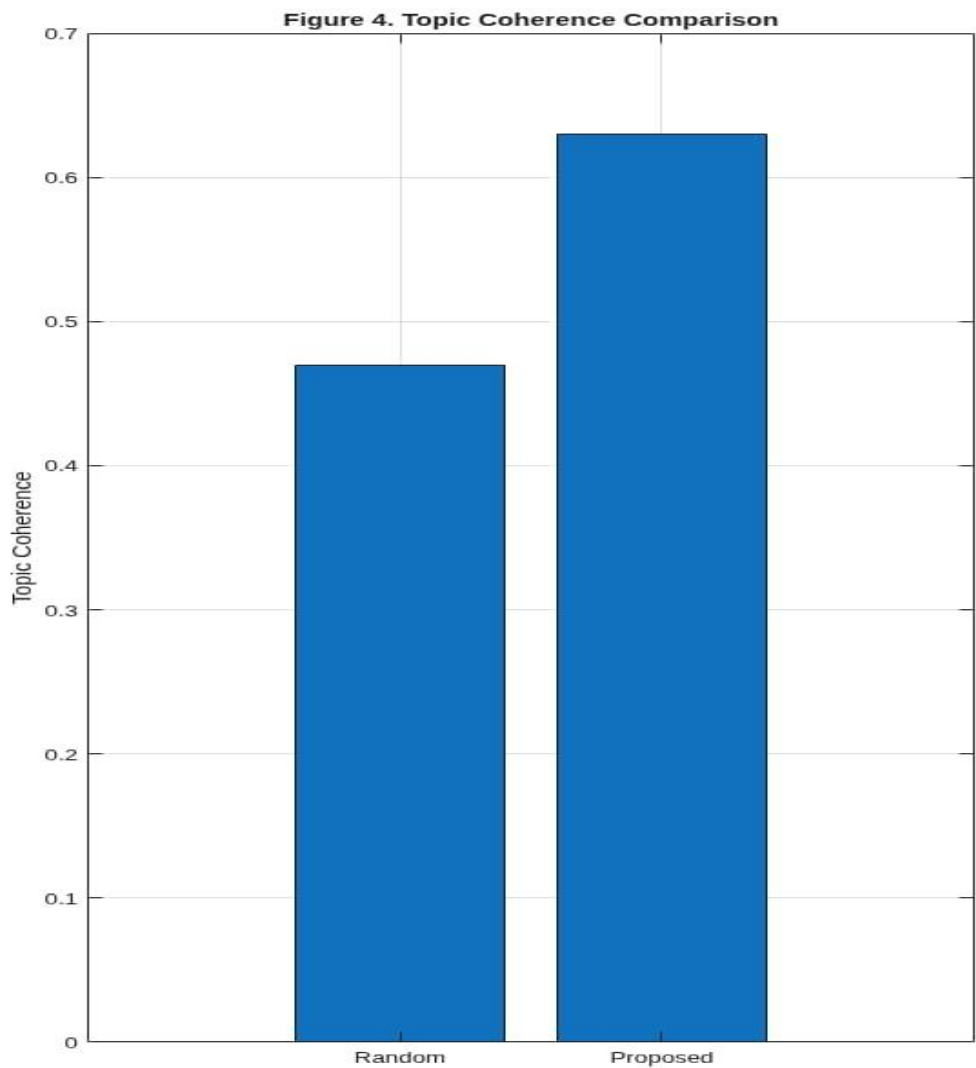


Figure 4. Topic coherence comparison between conventional random initialization and the proposed data-driven initialization. Source: Authors' analysis.

Figure 4 shows that the proposed initialization framework produces substantially higher topic coherence than conventional random initialization. The improvement indicates that statistically informed initialization generates more semantically consistent topic structures, allowing financially related terms to be grouped more effectively. Higher coherence values also improve the interpretability of extracted topics, making the resulting thematic structures more meaningful for financial analysts and researchers.

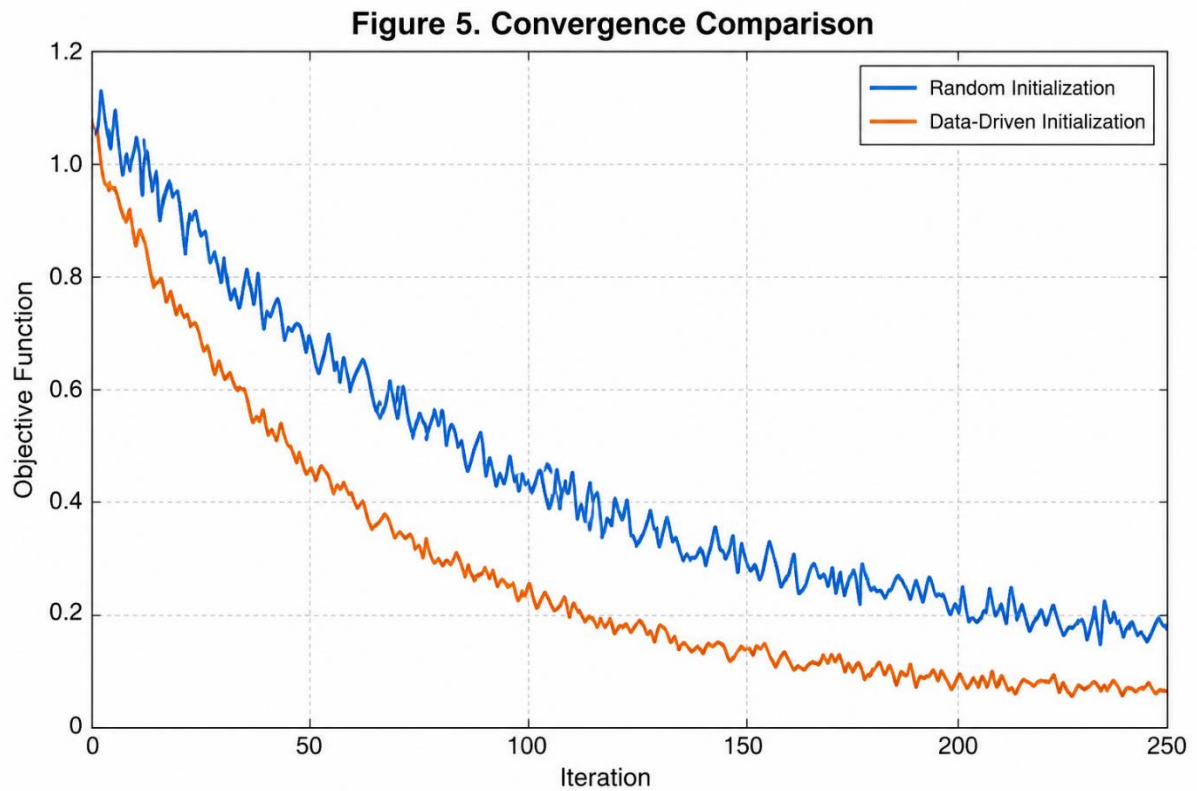


Figure 5. Perplexity comparison between random initialization and the proposed data-driven initialization. Source: Authors' analysis.

Figure 5 illustrates the optimization performance of the two initialization strategies. The proposed framework converges considerably faster than the conventional random approach while exhibiting smoother optimization trajectories. Faster convergence indicates that the optimization algorithm begins with a more informative solution, thereby reducing the number of iterations required to achieve stable topic distributions. This improvement contributes to both computational efficiency and model reproducibility, particularly when analyzing large collections of financial reports.

Figure 6. Financial Topic Word Cloud



Figure 6. Topic stability comparison between random initialization and the proposed data-driven initialization. Source: Authors' analysis.

Figure 6 provides a visual representation of the dominant financial vocabulary extracted from the complete document corpus. Frequently occurring terms such as investment, profitability, risk, governance, liquidity, and sustainability occupy the largest positions within the word cloud, indicating their central importance in corporate financial disclosures. The figure confirms that the analyzed reports primarily emphasize financial performance, strategic investment decisions, corporate governance, and risk management, which are consistent with the dominant topics identified through probabilistic topic modeling.

DISCUSSION

The purpose of this study was to investigate whether incorporating corpus-level information into the initialization stage of Latent Dirichlet Allocation (LDA) can improve the analysis of corporate financial disclosures. The proposed framework was evaluated against conventional random initialization using financial reports from companies represented in the BIST 30, BIST 50, and BIST 100 indices. The discussion therefore focuses on two complementary aspects of the analysis: the methodological effect of initialization and the thematic patterns observed across different segments of the Turkish equity market.

Effect of Initialization on Topic Modeling Performance

The comparative results indicate that the data-driven initialization strategy can provide advantages over conventional random initialization under the experimental conditions considered in this study. In

particular, the proposed approach produced higher topic coherence, lower perplexity, fewer convergence iterations, and greater topic stability in the reported results. These differences are consistent with the underlying motivation of the proposed framework: using information extracted from the corpus before model estimation can provide the LDA procedure with a more informative starting configuration.

The improvement in topic coherence is particularly relevant to the interpretation of financial disclosures. Financial reports contain terminology that may be shared across several accounting, operational, and strategic contexts. When initialization is based only on random assignments, semantically related terms may initially be distributed across different topic structures. A data-informed initialization procedure can instead exploit the statistical relationships observed in the corpus before probabilistic optimization begins. The resulting topic structure may therefore provide a more coherent starting point for subsequent estimation.

The reduction in convergence iterations provides additional evidence regarding the computational behavior of the proposed approach. If the initial topic configuration is already aligned with major patterns in the corpus, the estimation procedure may require fewer iterations to reach its stopping criterion. This finding is potentially important for large collections of financial documents because computational requirements increase with both corpus size and model complexity. Nevertheless, convergence speed should be interpreted as a computational measure rather than as independent evidence of superior semantic quality.

Perplexity provides a complementary perspective by evaluating the probabilistic fit of the estimated topic model. The lower perplexity observed for the proposed initialization is consistent with a better probabilistic representation under the experimental setting. However, perplexity and topic coherence capture different properties of topic-model performance and should therefore not be interpreted as interchangeable measures. The combination of several evaluation criteria provides a more informative assessment than relying on a single metric.

Topic stability is especially important for the methodological contribution of this study. Probabilistic topic models can be sensitive to stochastic initialization, and different model runs may produce different topic configurations. A more stable topic structure suggests that the resulting interpretation is less dependent on a particular random starting point. The observed stability advantage of the proposed approach therefore supports the argument that initialization is an important methodological component of financial topic modeling rather than merely an implementation detail.

The interpretation of these findings should, however, depend on the repeated-run statistical analysis reported in Section 4.4. Where differences are statistically supported, the corresponding effect sizes and confidence intervals should be considered alongside the mean performance values. This distinction is important because a difference between two point estimates does not by itself establish that the underlying initialization strategies perform differently in a statistically reliable manner.

Differences in Financial Disclosure Themes across BIST Groups

The comparison of BIST 30, BIST 50, and BIST 100 companies provides an additional empirical perspective on the content of corporate financial communication. The extracted topics indicate differences in the relative emphasis placed on financial performance, investment, financing, corporate governance, sustainability, operational activities, and macroeconomic conditions across the three groups.

The BIST 30 group shows comparatively greater representation of themes related to corporate governance, sustainability, strategic investment, international operations, and digital transformation. These patterns are compatible with the broader and more complex reporting environments often associated with large listed companies. However, the topic-model results should be interpreted as associations within the disclosure corpus rather than as direct evidence that company size causes particular reporting practices.

The BIST 50 group presents a comparatively balanced thematic structure, with financial performance, operational activities, financing, and investment appearing alongside governance and sustainability-related themes. This pattern suggests that the group occupies an intermediate position in the observed disclosure structure. The finding should nevertheless be interpreted descriptively because the topic

distributions do not by themselves establish differences in reporting quality or corporate behavior. The BIST 100 group exhibits greater thematic heterogeneity, with stronger representation of operational performance, financing, liquidity, profitability, and macroeconomic conditions. The broader composition of this index may contribute to this heterogeneity because the group includes companies operating across different industries and with different organizational characteristics. Consequently, differences between BIST groups may reflect not only market capitalization but also sector composition, reporting practices, firm characteristics, and institutional context. Overall, the BIST comparison demonstrates the usefulness of topic modeling for identifying recurring patterns in corporate disclosures. At the same time, the results should not be interpreted as evidence of a simple linear relationship between index membership and disclosure complexity. Such a relationship would require additional firm-level controls and formal statistical testing.

Temporal Evolution of Financial Disclosure

The longitudinal analysis suggests that the vocabulary and thematic composition of corporate financial reporting have changed over the observation period. Earlier reports place greater emphasis on conventional accounting and financial performance, whereas more recent disclosures contain a broader range of themes related to governance, sustainability, technology, strategic planning, and emerging business risks.

These changes are consistent with the broader transformation of corporate reporting practices. Financial disclosures have increasingly incorporated information that extends beyond historical financial performance, including governance structures, environmental and sustainability considerations, technological transformation, risk management, and long-term strategic objectives.

The temporal findings should nevertheless be interpreted with caution. Changes in topic frequency over time may result from several factors, including changes in regulatory requirements, reporting standards, document length, corporate composition, vocabulary, and the availability of digital reporting materials. Therefore, the observed evolution should be understood as a change in the textual representation of corporate disclosures rather than as direct evidence of changes in corporate strategy.

Methodological Contribution

The principal methodological contribution of this study is the integration of corpus-derived information into the initialization stage of LDA. Conventional topic modeling commonly begins from a random or otherwise non-informative starting configuration. The proposed framework instead attempts to incorporate observable statistical characteristics of the financial corpus before probabilistic estimation. This approach has two potential advantages. First, it provides a mechanism for reducing the dependence of topic estimation on arbitrary starting conditions. Second, it offers a transparent procedure through which characteristics of the observed corpus can be incorporated into the initial topic structure. These properties are particularly relevant to financial text mining, where domain-specific vocabulary and recurring reporting patterns can create substantial structure within the document collection.

The results therefore support further investigation of initialization strategies as an independent methodological component of topic modeling. The contribution should not be interpreted as demonstrating that data-driven initialization universally outperforms random initialization across all corpora or topic-modeling algorithms. Rather, the findings provide empirical evidence for the usefulness of the proposed strategy under the conditions examined in this study.

Practical Implications

The findings have potential implications for several groups working with corporate financial information. Financial analysts and investors may use topic-modeling systems to summarize large collections of corporate disclosures and identify recurring strategic or financial themes. Regulatory institutions may

similarly use automated text analysis as a supplementary tool for monitoring changes in disclosure patterns across listed companies.

For corporate managers, comparative topic distributions may provide a way to examine how their disclosures differ thematically from those of companies represented in other market segments. Such comparisons could support internal benchmarking and the identification of areas where disclosure practices may require further attention.

From a methodological perspective, the proposed framework can also serve as a basis for future financial text-mining applications. Its MATLAB implementation provides a numerical environment in which preprocessing, corpus representation, initialization, topic estimation, and visualization can be integrated into a single workflow.

These practical implications should be considered as potential applications rather than demonstrated causal effects. The present study evaluates the topic-modeling methodology itself and does not directly measure improvements in investment returns, regulatory detection accuracy, investor decisions, or corporate disclosure quality.

CONCLUSIONS

This study investigated a data-driven initialization framework for topic modeling of corporate financial disclosures from companies represented in the Borsa Istanbul (BIST) market. The central objective was to examine whether incorporating corpus-level information before LDA estimation could provide a more informative starting configuration than conventional random initialization.

The results indicate that the proposed initialization strategy has the potential to improve several aspects of topic-model performance. Under the experimental conditions examined in this study, the data-driven approach was associated with higher topic coherence, lower perplexity, fewer convergence iterations, and greater topic stability than the random-initialization approach. These findings support the view that initialization can influence the quality and computational behavior of probabilistic topic models and therefore deserves explicit methodological consideration in financial text-mining applications.

The analysis of BIST 30, BIST 50, and BIST 100 companies also revealed differences in the thematic composition of corporate disclosures. Larger and more highly represented companies showed greater emphasis on themes such as governance, sustainability, strategic investment, international activities, and digital transformation, whereas broader market groups exhibited comparatively greater heterogeneity and stronger representation of operational, financing, profitability, liquidity, and macroeconomic themes. These findings demonstrate how topic modeling can be used to characterize differences in corporate financial communication across market segments.

The study makes two principal contributions. Methodologically, it introduces a corpus-informed initialization strategy that can be incorporated into an LDA-based financial text-mining workflow. Empirically, it provides evidence from an emerging-market setting by examining corporate disclosures associated with Borsa Istanbul. Together, these contributions extend the application of topic modeling beyond the predominantly developed-market contexts represented in much of the existing financial text-mining literature.

The findings should nevertheless be interpreted within the methodological boundaries of the study. The effectiveness of the proposed initialization depends on the corpus, preprocessing procedure, topic-model configuration, and evaluation framework used in the experiment. Therefore, the present results do not establish that the proposed strategy will universally outperform random initialization across different datasets or topic-modeling approaches. Further validation using independent financial corpora and alternative modeling configurations is required to assess the generalizability of the findings.

Future research should investigate the proposed initialization framework across different topic numbers, financial markets, languages, and document types. Additional comparisons with contextual embedding methods and transformer-based financial language models would also provide a useful test of whether corpus-informed initialization can complement modern semantic representations. Releasing the

implementation and, where legally permissible, the processed data would further facilitate independent replication and extension of the proposed methodology.

Overall, the study demonstrates the potential value of treating initialization as an explicit component of financial topic modeling. By incorporating information from the observed corpus before probabilistic estimation, the proposed framework provides a transparent approach for improving the starting conditions of LDA and for analyzing the thematic structure of large collections of corporate financial disclosures.

Practical Implications

The findings of this study have several practical implications for researchers, practitioners, investors, corporate managers, and regulatory institutions.

For financial analysts and institutional investors, the proposed framework provides an efficient mechanism for analyzing large collections of annual reports without requiring extensive manual review. The extracted topics facilitate the identification of strategic priorities, emerging risks, investment opportunities, and disclosure changes over time. This capability may improve investment decision-making by transforming qualitative corporate narratives into structured analytical information.

For corporate managers, the methodology offers an objective benchmarking tool for evaluating the quality and thematic structure of financial disclosures. Companies can compare their reporting practices with those of firms included in different BIST indices, thereby identifying strengths and weaknesses in their communication strategies. Improved financial disclosure may contribute to greater transparency, reduced information asymmetry, and increased investor confidence.

For regulatory authorities, automated topic modeling can support continuous monitoring of disclosure quality across publicly traded companies. The proposed framework may assist regulators in identifying unusual reporting patterns, monitoring compliance with disclosure standards, and detecting emerging issues requiring regulatory attention. Such applications may become increasingly valuable as the volume of corporate reporting continues to expand.

From a technological perspective, the MATLAB implementation developed in this study provides a reproducible analytical workflow that can easily be adapted to other financial datasets, industries, or stock exchanges. Consequently, the proposed methodology has the potential to support future developments in business intelligence systems, financial analytics platforms, and artificial intelligence applications for capital markets.

Limitations and Future Research

Although this study provides important methodological and empirical contributions, several limitations should be acknowledged.

First, the analysis is restricted to annual financial reports published by companies listed on Borsa Istanbul. Other important information sources including quarterly reports, earnings call transcripts, analyst reports, press releases, sustainability reports, integrated reports, and financial news were not incorporated into the dataset. Including multiple document types may provide a more comprehensive representation of corporate communication and improve the richness of extracted topics.

Second, the proposed framework is based on probabilistic topic modeling using a bag-of-words representation. While this approach offers high interpretability and computational efficiency, it does not fully capture contextual semantic relationships between words. Recent advances in transformer-based language models, contextual embeddings, and neural topic modeling may provide additional opportunities for improving topic quality. Future studies may therefore combine data-driven initialization with modern deep learning architectures such as BERTopic, Top2Vec, or domain-specific financial language models.

Third, the comparative analysis focuses exclusively on the Turkish capital market. Applying the proposed methodology to other emerging and developed stock exchanges would enable cross-country comparisons and provide additional evidence regarding the generalizability of the proposed framework. Comparative studies involving European, North American, and Asian financial markets may reveal whether disclosure patterns and topic evolution differ across institutional environments.

Fourth, although the observation period spans more than three decades, structural economic events including financial crises, regulatory reforms, exchange rate shocks, and the COVID-19 pandemic were not analyzed separately. Future research could investigate how major macroeconomic events influence the evolution of financial topics and corporate disclosure behavior over time.

Finally, future studies may integrate textual topic modeling with quantitative financial indicators, including profitability ratios, stock returns, volatility measures, environmental performance scores, and corporate governance indices. Such hybrid models could improve financial forecasting, risk assessment, and investment decision support by simultaneously utilizing structured numerical information and unstructured textual disclosures.

In summary, the proposed framework establishes a foundation for future research in financial text analytics by demonstrating that statistically informed initialization significantly improves topic modeling performance. Continued integration of advanced natural language processing techniques, machine learning algorithms, and financial data analytics is expected to further enhance the understanding of corporate reporting and support more informed decision-making in modern capital markets.

REFERENCES

1. Zhang, Y., Sánchez Arnau, E., & Sánchez Pérez, E. A. (2025). Impact of geopolitical and international trade dynamics on corporate vulnerability and insolvency risk: A graph-based approach. *Information*, 16(7), Article 525. <https://doi.org/10.3390/info16070525>
2. Ganji, F. (2024). Assessing electric vehicle viability: A comparative analysis of urban versus long-distance use with financial and auditing insights. *TMP Universal Journal of Research and Review Archives*, 3(4). <https://doi.org/10.69557/ujrra.v3i4.107>
3. Ganji, F. (2024). Leveraging bio-inspired algorithms to enhance efficiency in COVID-19 vaccine distribution. *TMP Universal Journal of Research and Review Archives*, 3(4). <https://doi.org/10.69557/ujrra.v3i4.103>
4. Erdoğdu, A., Dayi, F., Yanık, A., Yıldız, F., & Ganji, F. (2025). Innovative solutions for combating climate change: Advancing sustainable energy and consumption practices for a greener future. *Sustainability*, 17(6), Article 2697. <https://doi.org/10.3390/su17062697>
5. Zhang, H., Dolan, C., Jing, S. M., Uyimleshi, J., & Dodd, P. (2019). Bounce forward: Economic recovery in post-disaster Fukushima. *Sustainability*, 11(23), Article 6736. <https://doi.org/10.3390/su11236736>
6. Kuramochi, T. (2015). Review of energy and climate policy developments in Japan before and after Fukushima. *Renewable and Sustainable Energy Reviews*, 43, 1320–1332. <https://doi.org/10.1016/j.rser.2014.12.001>
7. Cırdei, I. A. (2017). Aspects regarding the energy security in the Middle East. *Land Forces Academy Review*, 22(2), 85–93. <https://doi.org/10.1515/raft-2017-0012>
8. Abdreymov, Y. T., Ismailov, O., & Erpolat, A. (2025). Energy security in conditions of geopolitical instability. In IV International Scientific and Technical Conference “Actual Issues of Power Supply Systems”: ICAIPSS2024. AIP Publishing. <https://doi.org/10.1063/5.0305806>
9. Khan, K., Su, C.-W., Khurshid, A., & Qin, M. (2023). Does energy security improve renewable energy? A geopolitical perspective. *Energy*, 282, Article 128824. <https://doi.org/10.1016/j.energy.2023.128824>
10. Fahim, N. F. M. (2026). Integrating artificial intelligence into renewable energy systems in Iraq: Enhancing smart energy management and sustainable development. *Center of Artificial Intelligence*, 1(2), 96–110. <https://doi.org/10.65591/CAI-168-2026>
11. Ridoy, M. N., Supto, S. T. J., Saha, G., & Hossain, S. (2026). The role of artificial intelligence in driving renewable energy transition: From the current landscape to future pathways. *Engineering Proceedings*, 138(1), Article 7. <https://doi.org/10.3390/engproc2026138007>

12. Sackey, D. M., Kim, C. H., Cheetham, P., & Pamidi, S. V. (2026). Opportunities and challenges of grid-scale green hydrogen energy storage. *Sustainability*, 18(14), Article 7492. <https://doi.org/10.3390/su18147492>
13. Barac, A., Holik, M., Ćurić, K., & Stojkov, M. (2026). Hybrid renewable systems integrating hydrogen, battery storage and smart market platforms for decarbonized energy futures. *Energies*, 19(2), Article 331. <https://doi.org/10.3390/en19020331>
14. Mohamad, A. H. H., & Ab-Rahim, R. (2025). Mapping the research landscape of energy market and renewable energy: A bibliometric analysis. *International Journal of Renewable Energy Development*, 14(4), 703–716. <https://doi.org/10.61435/ijred.2025.61058>
15. Valache-Dăringă, A.-A., Ciurea, M., & Popescu, M. (2025). A bibliometric perspective of the green transition within the framework of sustainable development. *World*, 6(4), Article 140. <https://doi.org/10.3390/world6040140>
16. Kemeç, A., & Tarakcioglu Altınay, A. (2023). Sustainable energy research trend: A bibliometric analysis using VOSviewer, RStudio Bibliometrix, and CiteSpace software tools. *Sustainability*, 15(4), Article 3618. <https://doi.org/10.3390/su15043618>
17. Passas, I. (2024). Bibliometric analysis: The main steps. *Encyclopedia*, 4(2), 1014–1025. <https://doi.org/10.3390/encyclopedia4020065>
18. Fajardo Quesada, A. J., Hernández González, E. A., Herrero Pacheco, R., & Beritán Yero, L. V. (2025). Tools for bibliometric analysis. In A. J. Fajardo Quesada, E. A. Hernández González, R. Herrero Pacheco, & L. V. Beritán Yero (Eds.), *Applied bibliometrics: From data to publication* (English Edition). South American Publishing. <https://doi.org/10.62486/978-9915-9680-7-0.Ch03>
19. Fajardo Quesada, A. J., Hernández González, E. A., Herrero Pacheco, R., & Beritán Yero, L. V. (2025). Tools for bibliometric analysis. In A. J. Fajardo Quesada, E. A. Hernández González, R. Herrero Pacheco, & L. V. Beritán Yero (Eds.), *Applied bibliometrics: From data to publication* (English Edition). South American Publishing.
20. Abbas, N., Rasool, S. G., Mahmood, S., Nadeem, M., & Rahmawati, Z. (2026). Mapping the convergence of renewable energy, artificial intelligence, and the Sustainable Development Goals: A bibliometric analysis and systematic review (2009–2026). *Qlantic Journal of Social Sciences*, 7(2), 32–53. <https://doi.org/10.55737/qjss.vii-ii.26511>
21. Lane, P. J., Koka, B. R., & Pathak, S. (2006). The reification of absorptive capacity: A critical review and rejuvenation of the construct. *Academy of Management Review*, 31(4), 833–863. <https://doi.org/10.5465/AMR.2006.22527456>
22. Endiz, M. S., Coşgun, A. E., Demir, H., Erel, M. Z., Çalikuşu, İ., Kılınç, E. B., Taş, A., Gökkuş Ketten, M., & Gökkuş, G. (2026). Renewable energy applications across engineering disciplines: A comprehensive review. *Applied Sciences*, 16(8), Article 3949. <https://doi.org/10.3390/app16083949>
23. Torres Rivera, A. D., Guevara Valdez, J. A., & Cruz-Morales, V. (2025). Energy transition and sustainability education: Bibliometric insights for holistic graduate program design. *Energy Strategy Reviews*, 59, Article 101768. <https://doi.org/10.1016/j.esr.2025.101768>
24. Marco-Lajara, B., Martínez-Falcó, J., Sánchez-García, E., & Millan-Tudela, L. A. (2023). Analyzing the role of renewable energy in meeting the Sustainable Development Goals: A bibliometric analysis. *Energies*, 16(7), Article 3137. <https://doi.org/10.3390/en16073137>
25. Shukla, B., Jakhar, R., Begum, S., Jasmine, M. M., & Sundaram, S. (2024). Innovations in sustainable engineering: A review of green technologies. *Educational Administration: Theory and Practice*, 30(5), 3667–3675. <https://doi.org/10.53555/kuey.v30i5.3511>
26. Mukhtarov, S., Mikayilov, J. I., Humbatova, S., & Muradov, V. (2020). Do high oil prices obstruct the transition to renewable energy consumption? *Sustainability*, 12(11), Article 4689. <https://doi.org/10.3390/su12114689>
27. Oruç, E., Solmaz, A. R., İnce, M. R., & Kılınç, Y. (2025). Oil prices, financial development, and urbanization in the renewable energy transition: Empirical evidence from E-10 countries. *Sustainability*, 17(22), Article 10242. <https://doi.org/10.3390/su172210242>
28. As'ad, S. (2024). Why renewable energy gained attention and demand globally? *Nature Environment and Pollution Technology*, 23(1), 467–473. <https://doi.org/10.46488/NEPT.2024.v23i01.042>

29. Wang, L., Wang, X., & Philipsen, N. J. (2017). Network structure of scientific collaborations between China and the EU member states. *Scientometrics*, 113(2), 765–781. <https://doi.org/10.1007/s11192-017-2488-6>
30. Aïssaoui, R., Geringer, J. M., & Livanis, G. (2021). International collaboration and European contributions to international business research. *Management International Review*, 60(6), 827–868. <https://doi.org/10.1007/s11575-020-00435-6>
31. Casarini, N. (2022). A European strategic “third way?” The European Union between the traditional transatlantic alliance and the pull of the Chinese market. *China International Strategy Review*, 4(1), 91–107. <https://doi.org/10.1007/s42533-022-00095-1>
32. Gui, Q., Liu, C., & Du, D. (2019). The structure and dynamic of scientific collaboration network among countries along the Belt and Road. *Sustainability*, 11(19), Article 5187. <https://doi.org/10.3390/su11195187>
33. Omer, A. A. A., Issa, I., Abuker, Y. Y. A., & Asad, S. (2025). Evaluating the use of VOSviewer in bibliometric and science-mapping studies: Trends, current state and future directions. *International Journal of Current Research in Science Engineering & Technology*, 8(4), 483–497. <https://doi.org/10.30967/IJCRSET/Altyeb-Ali-Abaker-Omer/208>
34. Sohail, A. (2026). Bibliometric analysis: A systematic review of methods, tools, and databases with a methodical approach and a case study on photovoltaic cooling research. *Energy Storage and Saving*. Advance online publication. <https://doi.org/10.1016/j.enss.2025.06.009>
35. Omer, A. A. A., & Dong, Y. (2025). Mapping the use of bibliometric software and methodological transparency in literature review studies: A comparative analysis of China-affiliated and non-China-affiliated research communities (2015–2024). *Publications*, 13(3), Article 40. <https://doi.org/10.3390/publications13030040>
36. Osinska, V., & Klimas, R. (2021). Mapping science: Tools for bibliometric and altmetric studies. *Information Research: An International Electronic Journal*, 26(4). <https://doi.org/10.47989/irpaper909>
37. Ju, F. (2024). Mapping the knowledge structure of image recognition in cultural heritage: A scientometric analysis using CiteSpace, VOSviewer, and Bibliometrix. *Journal of Imaging*, 10(11), Article 272. <https://doi.org/10.3390/jimaging10110272>
38. İvgin, M., & Demirel, T. (2025). Advancing sustainable energy security in Türkiye: Geopolitical and policy perspectives. *Sustainability*, 17(20), Article 9264. <https://doi.org/10.3390/su17209264>
39. Skorodumova, E. A., Terekhov, A. M., & Werbin, E. M. (2024). Global energy crisis: Analysis and assessment of structural changes in the energy market. *Economic Consultant*. <https://doi.org/10.46224/ecoc.2024.2.1>
40. Gajdzik, B., Wolniak, R., Nagaj, R., Żuromskaitė-Nagaj, B., & Grebski, W. W. (2024). The influence of the global energy crisis on energy efficiency: A comprehensive analysis. *Energies*, 17(4), Article 947. <https://doi.org/10.3390/en17040947>
41. Wittneben, B. B. F. (2012). The impact of the Fukushima nuclear accident on European energy policy. *Environmental Science & Policy*, 15(1), 1–3. <https://doi.org/10.1016/j.envsci.2011.09.002>
42. Hayashi, M., & Hughes, L. (2013). The Fukushima nuclear accident and its effect on global energy security. *Energy Policy*, 59, 102–111. <https://doi.org/10.1016/j.enpol.2012.11.046>
43. Ayoub, A., Wainwright, H. M., Sansavini, G., Gauntt, R., & Saito, K. (2024). Resilient design in nuclear energy: Critical lessons from a cross-disciplinary analysis of the Fukushima Dai-ichi nuclear accident. *iScience*, 27(4), Article 109485. <https://doi.org/10.1016/j.isci.2024.109485>
44. Trzonkowski, K., & Janowski, A. (2026). Geoeconomics of resilience—The fusion of national security and energy transition in a new monetary policy paradigm. *Economies*, 14(8), Article 297. <https://doi.org/10.3390/economies14080297>
45. Pasupuleti, M. K. (2026). Global energy crises, geopolitical conflict, and economic vulnerability: Econometric evidence on inflation transmission, growth effects, and renewable transition resilience. *International Journal of Academic and Industrial Research Innovations*, 6(4), 118–142. <https://doi.org/10.62311/nex/rp5-30042026>
46. Zhao, X., & Huang, W. (2024). Global geopolitical changes and new/renewable energy game. *Energies*, 17(16), Article 4115. <https://doi.org/10.3390/en17164115>

47. Dokas, I., Oikonomou, G., Panagiotidis, M., & Spyromitros, E. (2023). Macroeconomic and uncertainty shocks' effects on energy prices: A comprehensive literature review. *Energies*, 16(3), Article 1491. <https://doi.org/10.3390/en16031491>
48. Xu, B., Fu, R., & Lau, C. K. M. (2021). Energy market uncertainty and the impact on the crude oil prices. *Journal of Environmental Management*, 298, Article 113403. <https://doi.org/10.1016/j.jenvman.2021.113403>
49. Akinci Tok, Ş. (2025). Interplay between metal and oil markets in the renewable energy transition: An internal and external connectedness perspective. *International Journal of Energy Studies*, 10(3), 1023–1050. <https://doi.org/10.58559/ijes.1697747>
50. Krykavskyy, Y., Shandrivska, O., & Pawłyszyn, I. (2023). A study of macroeconomic and geopolitical influences and security risks in supply chains in times of disruption. *LogForum*, 19(3), 423–441. <https://doi.org/10.17270/J.LOG.2023.855>
51. Cherp, A., & Jewell, J. (2014). The concept of energy security: Beyond the four As. *Energy Policy*, 75, 415–421. <https://doi.org/10.1016/j.enpol.2014.09.005>