Artificial Intelligence Applications And Innovations Vol. 1 No. 4 (2024)

Article history:
Received 24 May 2024
Revised 21 July 2024
Accepted 30 August 2024
Published online 01 October 2024

Scientific Trend Analysis of Artificial Intelligence Applications in Banking Models using Text Mining

Ramin Khoshchehreh Mohammadi [1], Mehrdad Hosseini Shakib [1*], Mahmood Khodam[1], Ali Ramezani[1]

[1] Department of Industrial Management, Ka.C., Islamic Azad University, Karaj, Iran

*** Corresponding author email address**:Mehrdad.shakib@kiau.ac.ir

Article Info

Article type:
Original Research

How to cite this article:
Khoshchehreh Mohammadi,R., Hosseini Shakib,M. Khodam.M. And Ramezani,A. (2025). Scientific Trend Analysis of Artificial Intelligence Applications in Banking Models using Text Mining. *Artificial Intelligence Applications and Innovations, 1*(4), 1-13.
https://doi.org/10.61838/jaiai.1.4.1

ABSTRACT

Reviewing scientific articles and comparing their status can identify scientific gaps and potential opportunities. This study focuses on the field of hybrid models of banking and artificial intelligence (AI). AI applications in banking have grown significantly, ranging from fraud detection and risk assessment to personalized customer services and automated trading systems. These technologies are not only enhancing operational efficiency but also transforming how financial institutions interact with their customers and manage risks. In this paper, after extracting data from the Scopus database, categorization was performed on 4,795 reputable articles over the past 14 years (2010-2023). Clusters were created using text mining techniques to assign subject labels in the interdisciplinary fields of AI and banking. The Box-Jenkins approach was then used to select a model on the data and predict and analyse trends over different periods. The results indicate the primary focus areas for applying AI in banking are: Innovation, Technologies and Digital Banking (58.89%), Commercial and Investment Banking (27.13%), Retail, Personal and Wealth Management Banking (9.49%), and International and Global Operations Banking (4.48%).

Keywords: *Banking models, Artificial Intelligence, Text mining, Classification, K-Nearest Neighbors, Box-Jenkins.*

1. Introduction

The banking industry is one of the most crucial economic sectors, handling vast amounts of financial and customer data. Given the importance of this data in strategic and operational decision-making, the use of data mining and Text Mining (TM) techniques in this industry is essential. According to a report by Chung et al. [1], banks that utilize advanced data analytics can increase their revenue by up to 25%. Furthermore, a study by Nguyen et al. [2] showed that the use of artificial intelligence and machine learning in banking can reduce operational costs by up to 22%. These statistics underscore the significance of data mining and text mining in the banking industry. The financial industry, like other industries, has been impacted by the revolution in data science and AI. TM techniques are crucial for knowledge discovery from databases, with banking databases being among the most significant types of such databases. An in-depth examination of successful methods for implementing AI in various variables and relationships within the banking industry globally can be highly important. Accordingly, Smith et al. [3] state that the future of banking is digital, and its impacts are highly influential upon close examination. According to Sarea et al. [4] and Gupta and Jain [5], the use of AI in the financial sector has transformed the industry in optimizing various operations, decision-making, transactions, and risk management. For further information on TM models, banking, and AI, readers can refer to Cheng and Sharmayne [6], Hassani et al. [7], Pattnaik et al. [8], and Moro et al. [9], and Doumpos et al. [10].

TM is a key technique in extracting valuable information from textual data, with widespread applications in the banking industry. This technique can be used in analyzing customer feedback, assessing credit risk, detecting fraud, and improving customer service. According to a Hristova [11], by 2025, over 75% of large banks will use text mining techniques to improve customer experience and risk management. Additionally, research conducted by IBM [12] shows that using text mining in analyzing customer complaints can reduce response time by up to 80%. The concept of TM, which refers to extracting all the needed information from huge amounts of textual data, has almost the same age as information retrieval itself. However, TM has unique and fundamental characteristics that distinguish it from information retrieval. TM helps in extracting useful information from textual data that is inherently unstructured, unorganized and irregular [13]. TM is an interdisciplinary field related to information science and knowledge discovery, computer science, mathematics, and computational linguistics [14]. It is also associated with information retrieval, data mining, machine learning, and statistics [15]. In the banking industry, text mining can be used to analyze financial reports, economic news, and legal documents to gain deeper insights into market trends, potential risks, and investment opportunities. This helps banks make more informed decisions and improve their performance. For further reading on the application of text mining in the banking industry, readers can refer to the comprehensive works of Pejić Bach et al. [16], Roeder and Palmer [17], Gupta et al. [18], Awotunde et al. [19], Nguyen and Huynh [20] and the recent survey by Plotnikova et al. [21].

According to Miner [22], TM helps in extracting useful information from textual data that is inherently unstructured, unorganized, and irregular. Hurst [23], one of the main pioneers in TM, provides a comprehensive definition of TM, in which he makes a clear fundamental distinction between TM and traditional information retrieval. Wei and Croft [24] demonstrated, based on the results of their studies, that topic modeling is not only beneficial but also outperforms many traditional approaches and clustering-based methods in information retrieval. Therefore, Zong et al. [25] conducted an investigation into researchers' needs for exploring and searching through the vast corpus of scientific texts, affirming TM as a structured and automated method for identifying the topics present within textual content - a task that traditional manual methods are incapable of accomplishing, yet which topic modeling is well-suited to perform. Accordingly, Saheb and Saheb [26] assert that this tool possesses significant potential for generating research ideas, encouraging collaboration among researchers, and more broadly, in the realm of scientific and research policymaking. Zengul et al. [27] state that recent advancements in TM and natural language processing have enabled a more comprehensive understanding of published articles to be attained. Natural language processing borrows a significant portion of its content from linguistics, computer science, and AI, facilitating interactions between computers and natural human language. TM encompasses semi-automated processes based on natural language processing, providing the capability for accurate and rapid analysis of large volumes of structured and unstructured data, as well as extracting patterns from texts by identifying keywords through their semantic features and co-occurrence within sentences. Regarding further TM research in various scientific domains, one can refer to Gagliardi and Albergo [28], Shojaee et al. [29], Woo et al. [30] Valtonen et al. [31], and Hickman et al. [32].

Advanced time series analysis and forecasting models, including the Box-Jenkins methodology, have greatly benefited the financial industry [33]. Box-Jenkins models, also known as Autoregressive Integrated Moving Average (ARIMA) models, are widely used for forecasting and understanding patterns in time series data, which is crucial in finance for applications such as stock price prediction, risk management, and portfolio optimization [34]. The Box-Jenkins approach involves identifying an appropriate ARIMA model by analyzing the stationarity and seasonality patterns in the data, estimating the model parameters, and then using the fitted model for forecasting future values [35]. This iterative process allows for capturing complex patterns and trends in financial time series, making Box-Jenkins models a powerful tool in the industry. The financial industry has also greatly benefited from advanced time series analysis and forecasting models, including the Box-Jenkins methodology. The Box-Jenkins approach involves identifying an appropriate ARIMA model by analyzing the stationarity and seasonality patterns in the data, estimating the model parameters, and then using the fitted model for forecasting future values. This iterative process allows for capturing complex patterns and trends in financial time series, making Box-Jenkins models a powerful tool in the industry. For further study in this area, readers can refer to Jafarian-Namin et al. [36], Mubarak [37], Mihalache and Bodislav [38], Yan et al. [39], and Qasim et al. [40].

Therefore, the main objective of this study is to examine the field of models of banking and AI by utilizing TM techniques to analyze a vast corpus of scientific articles, categorize them into relevant clusters, and predict future trends through time series analysis. In the following sections of this paper, Section 2 presents the methodology for a TM problem in the domain of banking and AI. Section 3 includes class analyses and the application of a predictive model. Finally, Section 4 offers conclusions and outlines approaches for future studies.

2. Problem-Solving Methodology

2.1. *Data Preparation*

For this research, the initial step in gathering information involved identifying reputable international journals within the field of banking and AI. To obtain a comprehensive list, various methods were employed, including expert opinions, electronic source searches using relevant keywords, and specialized databases related to banking and AI. This approach aimed to prevent overlooking any pertinent issues within this domain. These topics have been mainly categorized according to the common research classifications, as described in Table 1. Some of the keywords related to each class are listed in Table 1 to provide further insight into the respective class.

Table 1. Banking services and operations

Commercial and Investment Banking	Retail, Personal and Wealth Management Banking	Innovation, Technologies and Digital Banking	International and Global Operations Banking
AI-Financial Advisory Services	AI-Wealth Management	AI-Mobile Banking	AI-Foreign Exchange
AI-Asset Management	AI-Robo-Advisory	AI-Online Banking	AI-Global Payments
AI-Securities Trading	AI-Portfolio Management	AI-Blockchain Finance	AI-International Trade Finance
		AI-Fintech	

The classification presented covers key areas where AI is being applied across various banking services and operations. This categorization has been developed through a comprehensive evaluation and analysis of different use cases, industry trends, and emerging applications of AI in the banking sector. The categories have been organized based on the primary banking activities and functions they fall under, such as commercial and investment banking, retail and wealth management, digital banking innovations, and international operations. Within each category, specific AI applications and use cases have been identified, capturing the diverse ways AI is being leveraged to enhance efficiency, improve customer experiences, enable data-driven decision-making, and drive innovation in the banking industry. It's important to note that while this classification aims to be comprehensive, the field of AI in banking is rapidly evolving, with new applications and use cases emerging continually. Additionally, efforts have been made to consider the core domains and include relevant keywords and concepts closely related to each category. However, some AI applications may still span multiple categories or have cross-functional applications across different banking domains.

2.2. *The K-Nearest Neighbors Algorithm*

The KNN classification technique has been used to analyze the TM on the abstracts and to identify the clusters of the quality-related topics. This type of classification is widely used in text analysis and can be applied to any type of data with appropriate speed. In order not to ignore the contribution of repetitive variables in different documents than the variables that are more frequent in a particular document, the input of KNN classification information in RapidMiner software was considered as a file containing binary variables.

The KNN algorithm belongs to the supervised learning methods and is used for both classification and prediction. The functioning of this algorithm is based on observations and samples. In this method, the decision of which class a new sample belongs to is made by examining the K most similar neighboring samples. Among these K samples, the number of samples for each class is counted, and the new sample is assigned to the class to which the majority of neighbors belong. The first step in using KNN is to find a similarity measure, which is quantified by calculating the distance between data points. While this process is straightforward for numerical data, categorical variables require special treatment. Once the distances between different samples are measured, the set of previously classified samples can be used as a basis for classifying new samples. Let's assume we have data for class X and class Y in the form of equation (1), which are:

$$X = (x_1, x_2, \dots, x_n),\quad Y = (y_1, y_2, \dots, y_n) \tag{1}$$

The distance between two samples is calculated as the Euclidean distance using equation (2):

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

The KNN algorithm places a high computational burden on the computer because the computation time increases factorially with all points. In contrast, employing a decision tree or neural network for a new sample is a faster process. However, KNN requires a new computation for each new sample. Typically, to increase KNN's speed, all data are stored in memory. Understanding KNN models is straightforward when the number of predictor variables is small. This algorithm is also highly useful and efficient for constructing models for textual data, including handwriting recognition and satellite images, which involve various non-standard data types.

In KNN, generally, the larger the K, the less the effect of noisy data, and overall sensitivity to noisy data is eliminated. However, the classification boundaries become less distinct. Fortunately, this algorithm has also been successful in classification situations where the decision boundaries are highly irregular.

2.3. Text Processing and Classification

In text processing, processes such as format creation and text token cleaning are performed. During the text collection process, they may not be well organized. In this case, they are interpreted as lost information or irrational text integrity. If the texts are not examined properly, TM may lead to the phenomenon of generating incorrect output due to low-quality and incorrect input. In the preprocessing phase, documents are organized into a fixed number of predefined categories. Preprocessing ensures the successful implementation of text analysis, but it may consume significant processing time.

Given that this study's focus is on articles, specifically the words within article titles, abstracts, and keywords, each article title, abstract, and set of keywords were saved in a Microsoft Office Excel 2021 file for subsequent analyses. The articles were retrieved from the Scopus database. The English articles in the form of journal articles and conference papers were from reputable journals. Since the data collection for this research was conducted in the first half of 2024, the information for that year is incomplete; therefore, we analyzed the available data up to the end of 2023. Considering the large volume of information in the topics under review and the emphasis on new and emerging subjects, research from the last decade was examined. Thus, the database information covers the years 2010 to 2023. The content of each file is categorized based on the labels provided in

Table 1. Banking services and operations

. Therefore, the text processing process includes data collection, data analysis, data training and testing, model implementation, and result extraction. For this purpose, we considered 70 percent of the data for training and 30 percent for testing. Table 2 shows the relevant frequency and percentage for each category. Upon examining the number of articles in the field of AI in recent years, it can be concluded that this field has been one of the most flourishing research domains. In this regard, the studies in the form of journal and conference articles in the field of AI over a 14-year period from 2010 to 2023 indicate that AI, as a research and applied field, has garnered significant attention and has applications in many different domains, including manufacturing, healthcare, transportation, banking, and others. AI assists banking in making intelligent decisions and identifying financial patterns, enabling processes to be carried out more accurately and efficiently. Table 2 presents the frequency and percentage distribution of articles across four main classes related to the application of AI in the banking sector.

Table 2. Frequency and percentage of observations in each class

Class No.	Title of Class	Frequency	Percentage
1	Commercial and Investment Banking	1,301	27.13%
2	Retail, Personal and Wealth Management Banking	455	9.49%
3	Innovation, Technologies and Digital Banking	2,824	58.89%
4	International and Global Operations Banking	215	4.48%
Total observations		4,795	100%

The class with the highest frequency and percentage is "Innovation, Technologies and Digital Banking" with 2,824 articles, accounting for 58.89% of the total observations. This significant share highlights the strong focus and interest in leveraging AI for digital banking solutions, innovative technologies, and fintech applications within the banking industry. The second largest class is "Commercial and Investment Banking" with 1,301 articles, comprising 27.13% of the total observations. This class encompasses the use of AI in areas such as financial advisory services, asset management, securities trading, and corporate finance within the commercial and investment banking domains. The "Retail, Personal and Wealth Management Banking" class accounts for 9.49% of the total observations, with 455 articles. This class covers the application of AI in wealth management, robo-advisory services, portfolio management, and personal financial planning for retail and wealth management clients. The smallest class is "International and Global Operations Banking" with 215 articles, representing 4.48% of the total observations. This class focuses on the use of AI in areas such as foreign exchange, global payments, international trade finance, and cross-border transactions within the international banking operations.

The distribution of articles across these classes reflects the current trends and areas of focus within the banking industry, where digital transformation, innovation, and technological advancements are driving the adoption of AI solutions. However, the traditional banking domains of commercial, investment, retail, and global operations are also embracing AI to enhance efficiency, decision-making, and customer experience. The relatively smaller share of articles in the "Retail, Personal and Wealth Management Banking" and "International and Global Operations Banking" classes suggests that while these areas are exploring AI applications, the primary emphasis remains on digital innovation and commercial/investment banking activities. Nevertheless, as AI technologies continue to evolve and mature, their adoption across all banking domains is expected to grow, enabling streamlined processes, improved risk management, and personalized customer services. The overall trend of articles published in the field of AI applications in banking during the period of 2010-2023 is depicted in Figure 1.

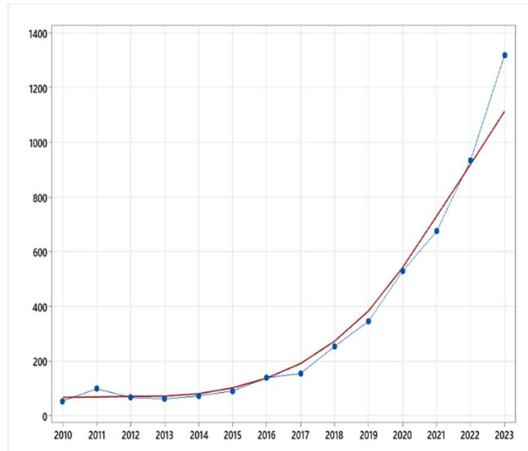
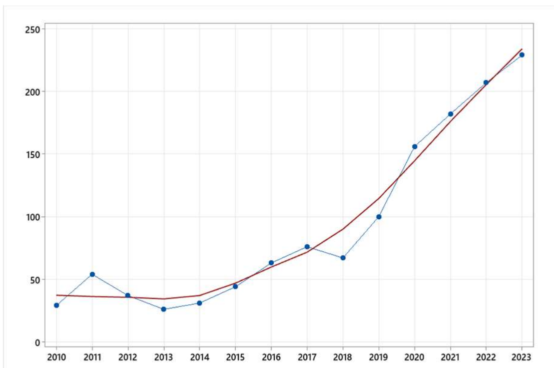


Figure 1. Overall trend of articles in the field of AI applications in banking (2010-2023)

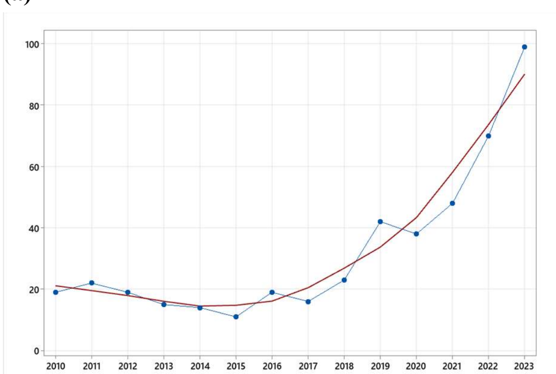
According to Figure 1, in 2010, the number of articles was relatively low at 54. However, this number started to gradually increase in the following years, reaching 99 in 2011 and 155 in 2017. A more substantial surge in the number of publications can be observed from 2018 onwards. The count nearly doubled from 254 in 2018 to 529 in 2020, indicating a growing interest and research activity in this domain. The upward trend continued in subsequent years, with the number of articles reaching 674 in 2021 and 934 in 2022. The data shows a remarkable peak of 1,318 articles published in 2023, which is more than double the count from the previous year. This consistent and substantial increase in the number of publications over the years suggests that the field of AI applications in banking has gained significant traction and attention from researchers, academics, and industry professionals. The rapid growth, especially in recent years, highlights the growing importance and adoption of AI technologies in the banking sector, potentially driven by factors such as technological advancements, digitalization efforts, and the need for efficient and innovative solutions in financial services.

Additionally, Figure 2 illustrates the separate trends for each of the specific banking domains covered in this study.

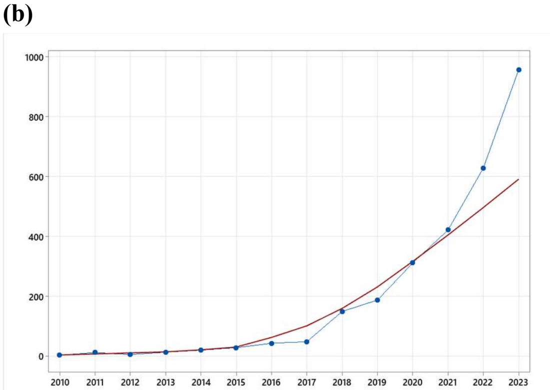
An analysis of Figure 2 reveals that the data highlights the increasing importance and adoption of AI technologies across various banking domains, with Innovation, Technologies and Digital Banking leading the way in terms of research and publication activity. The trends also suggest that domains like Commercial and Investment Banking, as well as Retail, Personal and Wealth Management Banking, have gained significant traction in recent years, while International and Global Operations Banking has seen a more gradual growth.



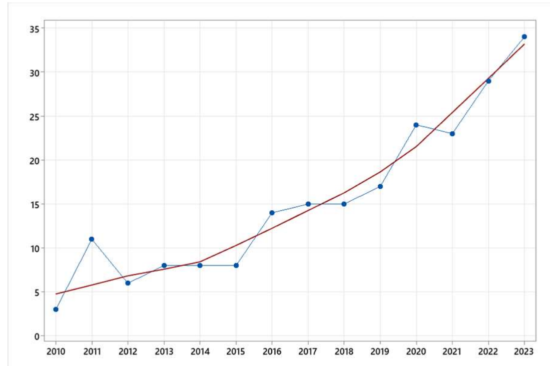
(a)



(b)



(c)



(d)

Figure 2. Separate trends for each banking domain covered in the study (2010-2023), (a) Commercial and Investment Banking, (b) Retail, Personal and Wealth Management Banking, (c) Innovation, Technologies and Digital Banking, (d) International and Global Operations Banking

3. Configuring a Forecasting Model

3.1. Box-Jenkins Model

Time series forecasting is a crucial task in various domains, and one prominent approach is the Box-Jenkins methodology, also known as the ARIMA modeling technique. This iterative process is particularly effective when dealing with autocorrelated data, where observations at different time points are correlated with one another. The Box-Jenkins methodology encompasses a systematic set of steps for identifying, estimating, and validating ARIMA models. It is a robust and versatile approach that can accommodate a wide range of time series patterns, including trends, seasonality, and irregularities. The iterative nature of the process allows for continuous refinement and optimization of the model to achieve accurate forecasts. The following subsections delineate the practical guidelines for implementing this iterative procedure.

3.1.1. Model Identification

In time series forecasting, the Box-Jenkins approach is widely used. The initial step is to assess the stationarity of the data and identify any significant seasonal patterns. This model identification phase involves detecting seasonality and determining the appropriate orders for the seasonal autoregressive and moving average components. The process begins with creating a time series plot to visually inspect patterns and stationarity. If the variance is unstable, a Box-Cox transformation may be applied to stabilize it.

If there is a shifting level or trend, regular differencing (d) can be performed to achieve stationarity in the mean. Once stationarity conditions are met, the focus shifts to identifying a tentative ARMA (Autoregressive Moving Average) model. Two graphical tools, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), are plotted to reveal the correlative structure. The ACF calculates the correlation between ordered pairs (Zt, Zt+k), while the PACF calculates the direct correlation between (Zt, Zt+k-1) without accounting for other time lags. By analyzing the ACF and PACF patterns, along with domain knowledge, a tentative ARMA model can be identified, including the orders for the autoregressive (AR) and moving average (MA) components. This tentative model serves as the starting point for subsequent parameter estimation and model validation stages in the Box-Jenkins methodology. For a more comprehensive understanding and mastery of the Box-Jenkins methodology, refer to Pankratz [35] in the field of time series analysis and forecasting.

3.1.2. Model Estimation

In the process of model specification, the parameter estimation is typically carried out during the second phase by employing the least squares method. According to the Pankratz [35], the crucial points pertaining to this phase are as follows: The estimated coefficients must adhere to the stationarity and invertibility conditions; the approximate absolute t-statistic value for each estimated coefficient should practically be around 2.0 or higher; if the estimated coefficients exhibit a high degree of correlation, i.e., an absolute correlation of 0.9 or greater, they may be deemed of poor quality, in which case, selecting an alternative adequate model with lower correlation is preferable; a model with a smaller adjusted root mean squared error (RMSE), which serves as an estimate of the random shocks' standard deviation, is desirable as it provides a better fit to the available data and tends to generate forecasts with smaller error variance. Another measure to assess the closeness of fit is the mean absolute percentage error (MAPE), which offers an indication of the expected accuracy of a forecasting model, described by the following expressions:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{a}_t}{Z_t}\right|.100 \tag{4}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\hat{a}_t^2} \tag{5}$$

The issue known as coefficient near-redundancy arises when the estimated coefficients of the non-seasonal AR and MA terms, both included in mixed models, have approximately similar values; in such cases, they should be removed to avoid non-parsimonious models with unstable estimated coefficients. Estimating parameters for Box-Jenkins models constitutes a complex non-linear estimation problem. Consequently, parameter estimation should be delegated to high-quality software programs capable of fitting Box-Jenkins models; fortunately, many commercial statistical software packages currently possess this capability.

3.1.3. Diagnostic Tests

For a Box-Jenkins model to be deemed adequate, the residuals must exhibit characteristics akin to white noise, essentially representing random drawings from a fixed distribution with constant mean and variance. When the Box-Jenkins model aligns well with the data, the residuals should conform to these assumptions. The calculation of residuals ($\hat{a}_t$), which serve as estimations of random errors, is performed using the following equation as per [35]:

$$\hat{a}_t = Z_t - \hat{Z}_t \tag{3}$$

Where Ŷt denotes the predicted value at time t, obtained from the previously established model. During the diagnostic checking phase, the assessment of the estimated model's statistical adequacy is conducted by scrutinizing the independence and normality of the residuals. In cases where the diagnosed model proves inadequate, a return to the identification stage becomes necessary to tentatively select an alternative model.

The evaluation of the model's validity commences with a visual inspection of the residuals plotted against time. Furthermore, the ACF and PACF plots for the residuals are examined, with statistical limits applied at a pre-defined confidence level. An overall assessment of the estimated residual autocorrelations is also undertaken through an approximate chi-squared statistic (Ljung-Box), defined as follows according to [41]:

$$\tilde{Q} = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}r_k^2(\hat{a}) \sim \chi^2(K - p - q) \quad (6)$$

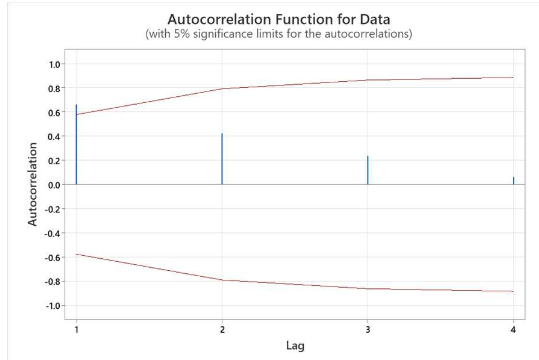
If the test result is significant, it signifies that the set of residuals lacks independence, thereby rendering the model inadequate. Consequently, a return to the identification phase is required to explore other tentative models. Additional checks can be performed by analyzing various plots. The normal probability plot and histogram are utilized to determine if the residuals' distribution approximates normality. The plot of residuals versus the fitted values facilitates the assessment of variance homogeneity over time, which should exhibit a random pattern. Once the estimated model is validated, implying that the residuals adhere to the white noise condition, an evaluation of the model's forecasting accuracy can be undertaken.

3.2. Model for Forecasting of Combined AI and Banking Articles

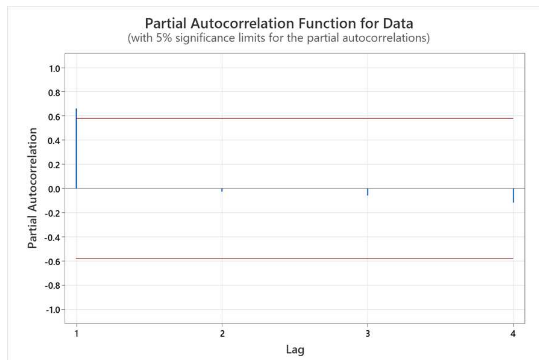
To explore upcoming trends in AI applications in banking, this section initially examines and forecasts the overall pattern of articles. Then, it delves into thematic classes within this domain. For data stability, we applied the Box-Cox transformation, specifically using the square root method. Afterward, we generated Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots for the transformed time series, as depicted in Figure 3. Interestingly, implementing first-order differencing rendered the time series stationary. Hence, we proceeded with further analysis using the ARIMA (1, 1, 1) model. Moreover, Table 3 presents a detailed analysis of the ACF and PACF results. Notably, the T value for the first lag in the ACF and PACF plot does not fall within the 95% confidence interval (±1.96), indicating a lack of significance. Therefore, due to the instability in the data mean, a first-order differencing is performed on the data.

Table 3. Analysis Results of ACF, and PACF

Lag	ACF	T	LBQ	PACF	T
1	0.661646	2.48	7.54	0.661646	2.48
2	0.422846	1.16	10.88	-0.026556	-0.10
3	0.236658	0.59	12.02	-0.058693	-0.22
4	0.061917	0.15	12.11	-0.116540	-0.44



(a) Autocorrelations



(b) Partial Autocorrelations

Figure 3. Illustrative ACF and PACF Charts of Time Series Data

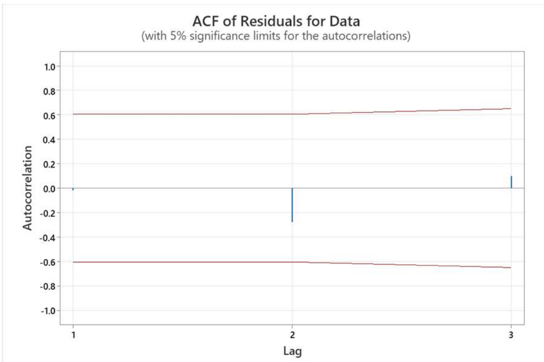
After fitting the ARIMA (1,1,1) model to the data, it was observed that the P-values for AR (1) and MA (1) were not statistically significant, indicating that these parameters did not have a significant impact on the model. Additionally, the residuals did not adhere to a normal distribution, suggesting that the model assumptions were not met. As a result of these findings, the ARIMA (1,1,1) model was deemed inadequate, prompting the exploration of alternative models. After thorough investigation, the ARIMA (1,1,2) model emerged as the most suitable choice. Notably, the distribution of residuals in this model demonstrated conformity to a normal distribution, validating its appropriateness for the dataset. Table 4 presents the detailed results obtained from the output of the Minitab software for the ARIMA (1,1,2) model, providing insights into the model's parameters and performance metrics.

Table 4. Performance Metrics of ARIMA (1,1,2) Model

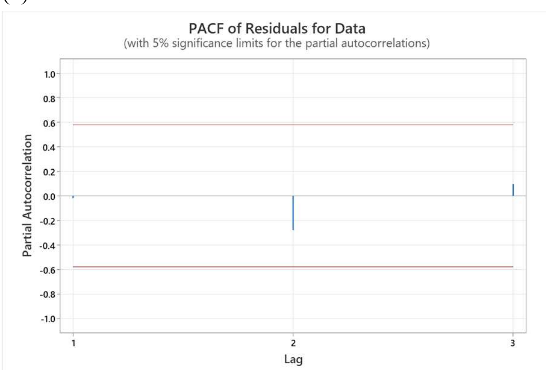
Type	Coef	E Coef	T-Value	P-Value
AR 1	1.008	0.217	4.65	0.001
MA 1	0.779	0.252	3.09	0.011

MA 2	-0.962	0.342	-2.81	0.018

The ACF and PACF plots of the residuals are presented in Figure 4 below. These plots are essential for assessing the adequacy of the model fit. From the plots, it is evident that the model has been appropriately fitted, as the patterns observed in the plots do not significantly deviate from the expected range, indicating satisfactory model performance.



(a) Autocorrelations



(b) Partial Autocorrelations

Figure 4. Illustrative ACF and PACF Charts of Residuals

Furthermore, Figure 5 displays additional diagnostic plots, which provide further insights into the model's performance. Additionally, the Anderson-Darling test, as depicted in Figure 6, is employed to assess the normality of the residuals. The obtained P-value from this test confirms that the distribution of residuals adheres to a normal distribution, strengthening the validity of the model.

Overall, the combination of diagnostic plots and statistical tests reaffirms the suitability of the ARIMA (1,1,2) model in capturing the underlying patterns in the data, providing a reliable basis for analysis and prediction.

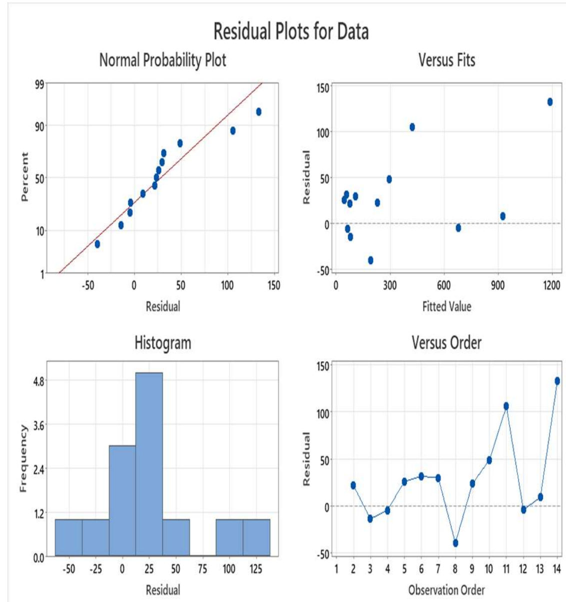


Figure 5. Residual Analysis Plots

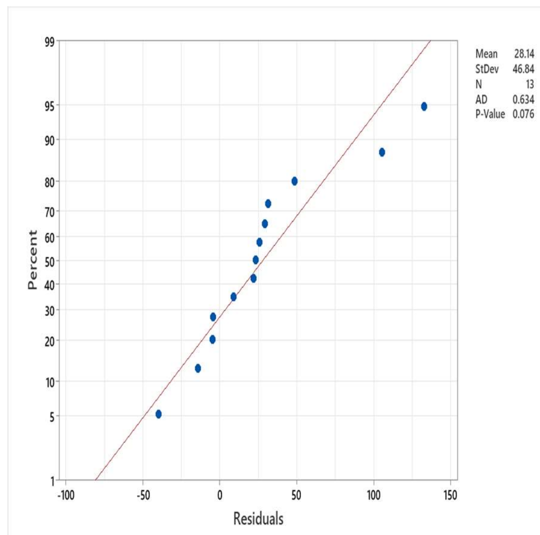


Figure 6. Probability Normal Plot of Residuals (Anderson-Darling Test)

Therefore, the forecasting trend for the next three years of articles in the field of AI and banking has been examined. This estimation is illustrated in Figure 7, while the predicted values and upper and lower bounds are presented in Table 5. It is evident from Figure 7 and Table 5 that predictions and upper and lower bounds have also been estimated for previous periods.

Consequently, considering that the number of published articles has fallen within the predicted range for these periods, it is anticipated that the trend of article publication, as depicted in Table 5, will continue from 2024 to 2026. For instance, in the year 2024, it might be expected that an average of 1425 articles will be published, with a minimum of 1081 and a maximum of 1768 articles.

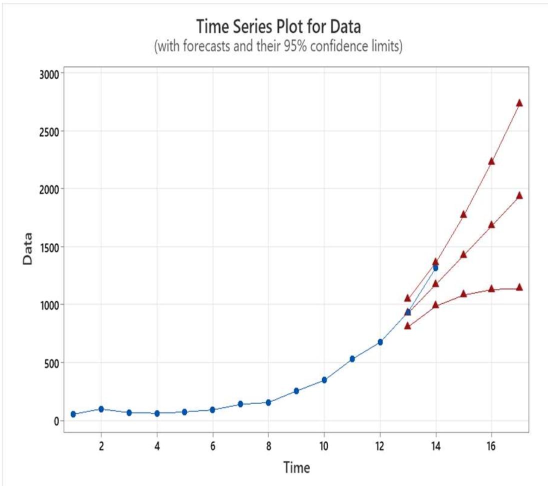


Figure 7. Forecasting Trend of Articles in the Field of AI and Banking (2024-2026)

Table 5. Predicted Values and Bounds for Articles in the Field of AI and Banking (2024-2026)

Period	Forecast	Lower	Upper	Actual
2022	925	806	1043	934
2023	1174	986	1362	1318
2024	1425	1081	1768	-
2025	1678	1128	2229	-
2026	1934	1138	2729	-

3.3. Model for Forecasting Labeled Classes

For predicting the labeled classes in the field of AI and banking in this study, considering the extensive computations, model selection, and validation process, it suffices to mention only the results. The prediction of data related to research topic categories was conducted using time series, as presented in Table 6. The validation of the obtained models was performed according to section 3.3.

Table 6. Predicted Labeled Classes in AI and Banking Research

Classes	Model	Period	Forecast	Lower	Upper
Commercial and Investment Banking	ARIMA (1,0,1)	2024	219	117	321
		2025	231	105	357
		2026	244	96	393
Retail, Personal and Wealth Management Banking	ARIMA (1,0,0)	2024	48	17	90
		2025	52	21	97
		2026	57	26	103
Innovation, Technologies and Digital Banking	ARIMA (1,3,0)	2024	1010	496	1523
		2025	1254	390	2118
		2026	1545	207	2884
International and Global Operations Banking	ARIMA (2,3,1)	2024	34	22	46
		2025	40	25	56
		2026	43	25	61

Table 6 summarizes predicted trends in various categories within AI and banking research over specified time periods. Each category is associated with a predictive model, indicating anticipated trajectories without specific numerical values. The predictive models employed include ARIMA (1,0,1), ARIMA (1,0,0), ARIMA (1,3,0), and ARIMA (2,3,1), reflecting diverse trends across commercial and investment banking, retail and personal banking, innovation and digital banking, as well as international and global operations banking. These predictions provide valuable insights into the expected research directions within the AI and banking sectors, facilitating informed decision-making processes. Stakeholders can utilize these projections to make informed decisions regarding resource allocation, research focus areas, strategic planning, and investment strategies within the AI and banking sectors.

4. Conclusion

This study aimed to analyze research trends in the field of AI applications in the banking industry by employing TM techniques on a large corpus of scientific articles from the Scopus database. A total of 4795 articles published between 2010 and 2023 were categorized into four main classes: (1) Commercial and Investment Banking, (2) Retail, Personal and Wealth Management Banking, (3) Innovation, Technologies and Digital Banking, and (4) International and Global Operations Banking. The results showed that the "Innovation, Technologies and Digital Banking" class had the highest frequency (58.89%), indicating a strong focus on leveraging AI for digital banking solutions, innovative technologies, and fintech applications within the banking industry. The "Commercial and Investment Banking" class followed with 27.13% of articles, covering AI applications in areas such as financial advisory services, asset management, and securities trading.

To forecast future trends, the Box-Jenkins (ARIMA) methodology was employed. The predictions for the number of articles in the field of AI and banking from 2024 to 2026 suggest a continued upward trend, with an expected average of 1425 articles in 2024, 1678 in 2025, and 1934 in 2026 (Table 5). Additionally, individual ARIMA models were developed to forecast the trends within each of the four classes (Table 6). The "Innovation, Technologies and Digital Banking" class is expected to maintain its dominance, with predicted upper bounds reaching 1523 articles in 2024, 2118 in 2025, and 2884 in 2026.

In conclusion, this study highlights the growing importance of AI applications in the banking industry, particularly in the areas of digital banking, innovation, and financial technologies. The analysis and forecasting provide valuable insights for stakeholders in the banking and AI sectors, enabling informed decision-making processes regarding research focus areas, resource allocation, and strategic planning. To harness the full potential of AI in the banking sector, future efforts should focus on developing advanced AI models for analyzing complex data and delivering personalized services. Integrating AI with emerging technologies like blockchain and IoT can create secure and efficient banking systems. Addressing ethical, legal, and security challenges, as well as training the workforce in AI, is crucial for successful adoption. Fostering collaboration between banks, researchers, and tech companies can drive joint research and innovation. Continuous monitoring, refining regulatory frameworks, and promoting public awareness will ensure responsible AI implementation, protect consumer rights, and build stakeholder trust. By embracing these recommendations, the banking industry can leverage AI's transformative power to enhance operations, mitigate risks, and provide superior customer experiences in the era of digital banking.

Authors' Contributions

All authors equally contributed to this study.

Declaration

None.

Transparency Statement

None.

Acknowledgments

None.

Declaration of Interest

The authors declare that they have no conflict of interest. The authors also declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

Not applicable.

References

[1] V. Chung, M. Gomes, S. Rane, S. Singh, and R. Thomas, "Reimagining customer engagement for the AI bank of the future," *McKinsey & Company,* 2020. [Online]. Available: https://www.mckinsey.com/industries/financial-services/our-insights/reimagining-customer-engagement-for-the-ai-bank-of-the-future.

[2] D. K. Nguyen, G. Sermpinis, and C. Stasinakis, "Big data, artificial intelligence and machine learning: A transformative symbiosis in favour of financial technology," *European Financial Management,* vol. 29, no. 2, pp. 517-548, 2023. [Online]. Available: https://doi.org/10.1111/eufm.12365.

[3] A. Smith and H. Nobanee, "Artificial intelligence: in banking A mini-review," *Available at SSRN 3539171,* 2020. [Online]. Available: https://doi.org/10.2139/ssrn.3539171.

[4] A. Sarea, M. R. Rabbani, M. S. Alam, and M. Atif, "Artificial intelligence (AI) applications in Islamic finance and banking sector," in *Artificial Intelligence and Islamic Finance*: Routledge, 2021, pp. 108-121.

[5] J. Gupta and S. Jain, "A study on Cooperative Banks in India with special reference to Lending Practices," *International Journal of Scientific and Research Publications,* vol. 2, no. 10, pp. 1-6, 2012. [Online]. Available: https://doi.org/10.15373/22778179/OCT2013/123.

[6] L. C. Cheng and L. R. Sharmayne, "Analysing digital banking reviews using text mining," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020: IEEE, pp. 914-918. [Online]. Available: https://doi.org/10.1109/ASONAM49781.2020.9381429.

[7] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data and Cognitive Computing,* vol. 4, no. 1, p. 1, 2020. [Online]. Available: https://doi.org/10.3390/bdcc4010001.

[8] D. Pattnaik, S. Ray, and R. Raman, "Applications of artificial intelligence and machine learning in the financial services

industry: A bibliometric review," *Heliyon,* 2024. [Online]. Available: https://doi.org/10.1016/j.heliyon.2023.e23492.
[9] S. Moro, P. Cortez, and P. Rita, "Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation," *Expert Systems with Applications,* vol. 42, no. 3, pp. 1314-1324, 2015. [Online]. Available: https://doi.org/10.1016/j.eswa.2014.09.024.
[10] M. Doumpos, C. Zopounidis, D. Gounopoulos, E. Platanakis, and W. Zhang, "Operational research and artificial intelligence methods in banking," *European Journal of Operational Research,* vol. 306, no. 1, pp. 1-16, 2023. [Online]. Available: https://doi.org/10.1016/j.ejor.2022.04.027.
[11] G. Hristova, "Text analytics for customer satisfaction prediction: A case study in the banking domain," in *AIP Conference Proceedings*, 2022, vol. 2505, no. 1: AIP Publishing. [Online]. Available: https://doi.org/10.1063/5.0100865.
[12] B. N. Silva, M. Diyan, and K. Han, "Big data analytics," *Deep learning: convergence to big data analytics,* vol. 13, p. 30, 2019. [Online]. Available: https://doi.org/10.1007/978-981-13-3459-7_2.
[13] D. Mittal and S. R. Agrawal, "Determining banking service attributes from online reviews: text mining and sentiment analysis," *International Journal of Bank Marketing,* vol. 40, no. 3, pp. 558-577, 2022. [Online]. Available: https://doi.org/10.1108/IJBM-08-2021-0380.
[14] T. Jo, *Text mining*. Springer, 2019.
[15] S. P. Nagarkar and R. Kumbhar, "Text mining: an analysis of research published under the subject category 'Information Science Library Science'in Web of Science Database during 1999-2013," *Library Review,* vol. 64, no. 3, pp. 248-262, 2015. [Online]. Available: https://doi.org/10.1108/LR-08-2014-0091.
[16] M. Pejić Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustainability,* vol. 11, no. 5, p. 1277, 2019. [Online]. Available: https://doi.org/10.3390/su11051277.
[17] J. Roeder and M. Palmer, "Document Representation for Text Analytics in Finance," in *Enterprise Applications, Markets and Services in the Finance Industry: 9th International Workshop, FinanceCom 2018, Manchester, UK, June 22, 2018, Revised Papers 9*, 2019: Springer, pp. 131-145. [Online]. Available: https://doi.org/10.1007/978-3-030-19037-8_9.
[18] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, "Comprehensive review of text-mining applications in finance," *Financial Innovation,* vol. 6, pp. 1-25, 2020. [Online]. Available: https://doi.org/10.1186/s40854-020-00205-1.
[19] J. B. Awotunde, E. A. Adeniyi, R. O. Ogundokun, and F. E. Ayo, "Application of Big Data with Fintech in Financial Services," in *Fintech with Artificial Intelligence, Big Data, and Blockchain*, P. M. S. Choi and S. H. Huang Eds. Singapore: Springer Singapore, 2021, pp. 107-132.
[20] B.-H. Nguyen and V.-N. Huynh, "Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach," *Journal of the Operational Research Society,* vol. 73, no. 1, pp. 102-121, 2022. [Online]. Available: https://doi.org/10.1080/01605682.2020.1784049.
[21] V. Plotnikova, M. Dumas, and F. P. Milani, "Data mining methodologies in the banking domain: A systematic literature review," in *Perspectives in Business Informatics Research: 18th International Conference, BIR 2019, Katowice, Poland, September 23–25, 2019, Proceedings 18*, 2019: Springer, pp. 104-118. [Online]. Available: https://lepo.it.da.ut.ee/~dumas/pubs/bir2019.pdf.
[22] G. Miner, *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.
[23] K. N. Singh, S. D. Devi, H. M. Devi, and A. K. Mahanta, "A novel approach for dimension reduction using word embedding: An enhanced text classification approach," *International Journal of Information Management Data Insights,* vol. 2, no. 1, p. 100061, 2022. [Online]. Available: https://doi.org/10.1016/j.jjimei.2022.100061.
[24] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 178-185. [Online]. Available: https://doi.org/10.1145/1148170.1148204.
[25] C. Zong, R. Xia, and J. Zhang, *Text data mining*. Springer, 2021.
[26] T. Saheb and M. Saheb, "Analyzing and visualizing knowledge structures of health informatics from 1974 to 2018: a bibliometric and social network analysis," *Healthcare informatics research,* vol. 25, no. 2, pp. 61-72, 2019. [Online]. Available: https://doi.org/10.4258/hir.2019.25.2.61.
[27] F. D. Zengul, N. Oner, J. D. Byrd, and A. Savage, "Revealing research themes and trends in 30 Top-ranking accounting journals: a text-mining approach," *Abacus,* vol. 57, no. 3, pp. 468-501, 2021. [Online]. Available: https://doi.org/10.1111/abac.12214.
[28] A. R. Gagliardi and F. Albergo, "Reducing Medical Errors via Data Mining Techniques: A Structured Literature Review and Future Research Agenda," in *The International Research & Innovation Forum*, 2023: Springer, pp. 723-736. [Online]. Available: https://doi.org/10.1007/978-3-031-44721-1_55.
[29] M. Shojaee, D. M. Imani, S. Jafarian-Namin, and A. Haeri, "Text mining, clustering, and forecasting horizons ahead in the field of quality and productivity," *International Journal of Productivity and Quality Management,* vol. 37, no. 4, pp. 559-577, 2022. [Online]. Available: https://doi.org/10.1504/IJPQM.2022.127508.
[30] S. E. Woo, L. Tay, and F. Oswald, "Artificial intelligence, machine learning, and big data: Improvements to the science of people at work and applications to practice," ed: Wiley Online Library, 2024.
[31] L. Valtonen, S. J. Mäkinen, and J. Kirjavainen, "Advancing reproducibility and accountability of unsupervised machine learning in text mining: Importance of transparency in reporting preprocessing and algorithm selection," *Organizational Research Methods,* vol. 27, no. 1, pp. 88-113, 2024. [Online]. Available: https://doi.org/10.1177/10944281221124947.
[32] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations," *Organizational Research Methods,* vol. 25, no. 1, pp. 114-146, 2022. [Online]. Available: https://doi.org/10.1177/1094428120971683.
[33] M. Susruth, "Financial forecasting: An empirical study on box–Jenkins methodology with reference to the Indian stock market," *Pacific Business Review International,* vol. 10, no. 2, pp. 115-123, 2017. [Online]. Available: https://www.scribd.com/document/527386792/1.
[34] B. Singh *et al.*, "Auto-regressive integrated moving average threshold influence techniques for stock data analysis," *International Journal of Advanced Computer Science and Applications,* vol. 14, no. 6, 2023. [Online]. Available: https://doi.org/10.1201/9781003369066.

[35] A. Pankratz, *Forecasting with univariate Box-Jenkins models: Concepts and cases*. John Wiley & Sons, 2009.
[36] S. Jafarian-Namin, S. M. T. Fatemi Ghomi, M. Shojaie, and S. Shavvalpour, "Annual forecasting of inflation rate in Iran: Autoregressive integrated moving average modeling approach," *Engineering Reports,* vol. 3, no. 4, p. e12344, 2021. [Online]. Available: https://doi.org/10.1002/eng2.12344.
[37] M. S. Mubarak, "Exploring The Forecasting Inflation in Kota Palu: An Application of the ARIMA Model," *Jurnal Ekonomi Pembangunan,* vol. 21, no. 01, pp. 1-14, 2023. [Online]. Available: https://doi.org/10.22219/jep.v21i01.25919.
[38] R.-P. MIHALACHE and D. A. Bodislav, "Forecasting the Romanian inflation rate: An Autoregressive Integrated Moving-Average (ARIMA) approach," *Theoretical & Applied Economics,* vol. 30, no. 1, 2023. [Online]. Available: https://store.ectap.ro/articole/1641.pdf.
[39] K. Yan, R. Gupta, and S. Haddad, "Statistical Analysis Dow Jones Stock Index—Cumulative Return Gap and Finite Difference Method," *Journal of Risk and Financial Management,* vol. 15, no. 2, p. 89, 2022. [Online]. Available: https://doi.org/10.3390/jrfm15020089.
[40] T. B. Qasim, H. Ali, N. Malik, and M. Liaquat, "Forecasting Inflation Applying ARIMA Model with GARCH Innovation: The Case of Pakistan," *Journal of Accounting and Finance in Emerging Economies,* vol. 7, no. 2, pp. 313-324, 2021. [Online]. Available: https://doi.org/10.26710/jafee.v7i2.1681.
[41] T. C. Mills, *Applied time series analysis: A practical guide to modeling and forecasting*. Academic press, 2019.