# CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

|  |  |
| --- | --- |
| Module Title: | Advanced Data Analytics  Big Data Storage and Processing |
| Assessment Title: | Forecasting Merchandise Trade Values between Ireland and International Partners Using Recurrent Neural Networks: A Time Series Analysis |
| Lecturer Name: | David McQuaid and Muhammad Iqbal |
| Student Full Name: | Fernando Tupa Leniz |
| Student Number: | 2023341 |
| Assessment Due Date: | 07/04/2024 |
| Date of Submission: | 07/04/2024 |

**Github link:** <https://github.com/FernandoDataAnalitycs/CA_sec_semester_1>

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Forecasting Merchandise Trade Values between Ireland and International Partners Using Recurrent Neural Networks: A Time Series Analysis

*Fernando Tupa Leniz*   
*Msc in Data analytics  
CCT college  
Dublin, Ireland  
2023341@student.ie*

*CCT college*

*Dublin, Ireland*

*2023341@student.ie*

# ABSTRACT

In the dynamic landscape of global economics, accurate prediction of merchandise trade values holds paramount importance for policymakers, businesses, and economists alike. This research paper investigates the efficacy of two prominent recurrent neural network architectures, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), in forecasting the value of merchandise trade between Ireland and various countries. Leveraging a comprehensive time series dataset, spanning multiple years, the study employs LSTM and GRU models to predict trade values and compares their effectiveness in capturing the intricate patterns inherent in international trade dynamics. Through rigorous evaluation and comparative analysis, we reveal insights into the performance differences between LSTM and GRU models, shedding light on their respective strengths and weaknesses in the context of merchandise trade prediction. Our findings not only contribute to advancing the methodology of time series forecasting but also underscore the \*\*vital importance of accurately predicting the value of merchandise trade\*\* for informed decision-making in the global economy.

# KEYWORDS

Merchandise trade;Time series prediction ;Recurrent Neural Network; LSTM;GRU; Comparative analysis; apache spark;

Hadoop

# Introduction

The globalization of economies has catalyzed an unprecedented expansion in international trade, transforming it into a cornerstone of contemporary economic activity. The intricacies of merchandise trade, encompassing the exchange of goods and services across national borders, underscore its vital role in shaping the economic landscape of nations worldwide [1]. As economies become increasingly intertwined, accurate forecasting of merchandise trade values emerges as a critical imperative for policymakers, businesses, and economists alike. The ability to anticipate fluctuations in trade dynamics facilitates informed decision-making, aids in the formulation of effective trade policies, and enables businesses to adapt strategies to changing market conditions [2].

In this era of data-driven decision-making, the application of advanced quantitative research methodologies has become increasingly prevalent in analyzing and predicting complex economic phenomena [3]. Machine learning techniques, in particular, have gained prominence for their ability to extract meaningful insights from large-scale datasets and model intricate patterns inherent in time series data. Recurrent neural networks (RNNs), a class of artificial neural networks designed to analyze sequential data, have demonstrated remarkable efficacy in time series forecasting tasks [4]. Among the variants of RNNs, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures have emerged as powerful tools for capturing long-range dependencies and handling temporal dynamics, making them well-suited for predicting the dynamics of merchandise trade [5].

Against this backdrop, this research paper aims to investigate the effectiveness of LSTM and GRU recurrent neural networks in forecasting the value of merchandise trade between Ireland and a diverse set of trading partners. By leveraging a quantitative research approach, specifically utilizing time series analysis and machine learning techniques, this study seeks to provide empirical evidence on the relative performance of different models for predicting the dynamics of international trade. Through a comprehensive evaluation and comparative analysis of LSTM and GRU models, this research endeavors to elucidate the strengths and limitations of each model in capturing the nuances of merchandise trade dynamics. Furthermore, the study aims to contribute to the advancement of quantitative methodologies in the field of economic forecasting while providing valuable insights for policymakers and stakeholders involved in international trade [6].

I.I Objective Statement:

The primary objective of this study is to assess the comparative performance of LSTM and GRU recurrent neural networks in forecasting merchandise trade values, with a focus on the trade relationships involving Ireland and various countries.

I.II Research Question:

The central research question guiding this investigation is: "What is the relative effectiveness of LSTM and GRU recurrent neural networks in forecasting the value of merchandise trade between Ireland and its trading partners?"

# METHODOLOGY

## II.I Research Location

Ireland, positioned in Northwestern Europe, offers a compelling research location for investigating the dynamics of international trade. With its transition from an agrarian-based economy to a thriving hub for technology, pharmaceuticals, and financial services, Ireland boasts a robust export-driven economy driven by advantageous tax policies and a skilled workforce. Extensive trade relations with countries such as the United States, Great Britain, and other European Union members encompass diverse sectors like machinery, electronics, and pharmaceuticals. However, challenges such as Brexit-related uncertainties and global trade tensions provide opportunities for innovative research endeavors aimed at enhancing trade forecasting models and policy frameworks. Research in this area not only sheds light on global economic interconnectedness but also informs policy decisions and fosters sustainable development initiatives, making Ireland an intriguing focal point for studying the value of merchandise trade.

## II.II Data collection

Data collection is a pivotal aspect in any research endeavor, particularly in the context of handling large-scale datasets such as those prevalent in international trade analysis. Apache Spark and Hadoop emerge as indispensable tools in managing and processing such voluminous datasets efficiently. Spark, with its distributed computing framework, offers significant advantages in terms of speed and scalability, making it well-suited for handling big data applications [7]. On the other hand, Hadoop provides a reliable and fault-tolerant distributed storage system, ensuring the resilience and integrity of data storage [8].

In the realm of international trade analysis, where datasets often encompass diverse variables and span extensive time periods, the capability to handle big data becomes paramount. Utilizing Apache Spark and Hadoop facilitates seamless data processing and analysis, enabling researchers to extract meaningful insights from massive datasets efficiently. By leveraging the parallel processing capabilities of Spark and the robust storage infrastructure of Hadoop, researchers can overcome the challenges posed by the sheer volume and complexity of trade data [9].

Moreover, the integration of Apache Spark and Hadoop in the data collection pipeline offers additional benefits such as fault tolerance, data redundancy, and scalability, which are crucial for ensuring the reliability and robustness of the analysis. The fault-tolerant nature of Hadoop's distributed file system (HDFS) ensures data integrity even in the event of hardware failures or system crashes [10]. Additionally, the scalability of both Spark and Hadoop allows researchers to seamlessly accommodate growing datasets and computational demands, thus future-proofing the data collection infrastructure [11].

In acquiring the dataset for international trade analysis, it is crucial to obtain reliable and authoritative sources of data. A notable platform for accessing datasets relevant to Ireland's trade activities is https://data.gov.ie. This platform hosts a wide array of datasets provided by various governmental agencies, offering researchers access to comprehensive and up-to-date information on Ireland's trade dynamics. Additionally, it is essential to adhere to licensing terms and regulations governing the usage of such datasets. The dataset obtained from https://data.gov.ie is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), as specified by the licensing information available at https://creativecommons.org/licenses/by/4.0/. Adhering to the terms of this license ensures that researchers can responsibly utilize the dataset for analysis and research purposes while acknowledging the source appropriately. This compliance with licensing terms underscores the ethical and legal considerations inherent in data collection and usage within the research framework.

## II.III Data preparation

Data preparation is a critical step in preparing the dataset for analysis, particularly in the context of large-scale trade data. Leveraging PySpark, we streamlined the preprocessing pipeline to ensure data integrity and suitability for time series forecasting. Initially, we addressed missing values, outliers, and inconsistencies to enhance the dataset's reliability. Following this, feature selection was meticulously executed, identifying the target variable, "VALUE," which represents the value of merchandise trade, and relevant predictor variables, including trade partners and specific trade categories.

To facilitate time series analysis, the dataset underwent transformation into a suitable format, enabling effective modeling of trade dynamics over time. Additionally, to promote stable model convergence, numerical features underwent normalization using standard scaling techniques. Finally, to assess model performance accurately, the dataset was partitioned into distinct training and testing sets, ensuring robust evaluation of LSTM and GRU models.

The selection of features in our study was guided by the aim of predicting the value of merchandise trade between Ireland and various countries, with a specific focus on the category "Chemical materials and products, n.e.s. (59)" in the VALUE column. This choice is substantiated by the significance of chemical trade within the global market, spanning diverse industries pivotal to economic growth. The broad scope of this category encapsulates miscellaneous chemical commodities, reflecting the multifaceted nature of international trade in chemicals.

Furthermore, our analysis zeroes in on specific countries, such as the USA and Great Britain, justified by their substantial roles as key trading partners of Ireland. These countries are pivotal destinations for Irish exports and primary sources of imports, underscoring their economic importance. By delving into trade relations with these influential partners, our research aims to uncover nuanced insights into Ireland's trade dynamics, offering valuable perspectives for economic analysis and policy formulation. Together, the selected features enable a targeted exploration of trends in chemical trade, shedding light on the intricate interplay between Ireland and its prominent trade counterparts.

II.IV Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are a class of neural networks particularly well-suited for sequential data processing, such as time series analysis, natural language processing, and speech recognition. Unlike traditional feedforward neural networks, RNNs have connections that form directed cycles, allowing them to exhibit dynamic temporal behavior. The key feature of RNNs is their ability to maintain a state or memory of previous inputs through recurrent connections.

Mathematically, an RNN processes input sequences over time steps to produce output sequences, where each output depends not only on the current input but also on previous inputs through hidden states . The computation at each time step can be represented by the following equations:

(1)

(2)

Here, *f* and *g* are activation functions, and are weight matrices, *bh* and *by* are bias vectors, *xt* is the input at time step *t,* ht is the hidden state at time step *t*, and *yt* is the output at time step *t*.

This recurrent connectivity enables RNNs to capture temporal dependencies in the data, making them suitable for tasks where the order of inputs matters. However, traditional RNNs suffer from the vanishing gradient problem, which can hinder their ability to learn long-range dependencies in sequential data.

II.V Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a specialized type of recurrent neural network (RNN) architecture designed to address the vanishing gradient problem commonly encountered in traditional RNNs. LSTMs introduce additional components called memory cells and gates to selectively store and access information over time. These gates, controlled by sigmoid activation functions, modulate the flow of information through the cell state, allowing LSTM networks to learn when to remember or forget information.

In an LSTM cell, there are three main gates: the forget gate, the input gate, and the output gate. The forget gate decides what information should be discarded from the cell state, the input gate decides what new information should be stored in the cell state, and the output gate decides what information should be output from the cell. By carefully regulating the flow of information through these gates, LSTMs are capable of capturing and preserving long-term dependencies in sequential data, making them well-suited for tasks such as natural language processing, time series prediction, and speech recognition.

II.VI Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) is another variant of recurrent neural network (RNN) architecture that addresses the challenges of learning long-range dependencies while being computationally more efficient than LSTMs. Unlike LSTMs, GRUs merge the cell state and hidden state into a single state vector, simplifying the architecture and reducing the number of parameters.

In a GRU cell, there are two main gates: the update gate and the reset gate. The update gate controls how much of the past information needs to be passed along to the future, while the reset gate determines how much of the past information to forget. By adaptively updating the state vector based on these gates, GRUs are able to capture temporal dependencies in sequential data while maintaining a simpler and more efficient architecture compared to LSTMs.

Both LSTMs and GRUs have proven to be effective in modeling sequential data and have been widely used in various applications, including machine translation, sentiment analysis, and time series forecasting. The choice between them often depends on factors such as the complexity of the task, the amount of available training data, and computational resources. While LSTMs are known for their ability to capture long-term dependencies, GRUs offer a simpler design with comparable performance in many applications, making them a popular choice for practitioners seeking a balance between model complexity and efficiency.

III. MEASURE OF ACCURACY

Mean Squared Error (MSE) is a fundamental metric used to evaluate the performance of predictive models, particularly in regression tasks like time series forecasting. It quantifies the average squared difference between the predicted values and the actual values in a dataset. MSE is widely utilized due to its simplicity and effectiveness in providing a comprehensive measure of prediction accuracy.

The equation for MSE is expressed as follows:

(3)

Where:

- *n* represents the total number of data points in the dataset.

- *yi* denotes the actual value of the target variable for the *i*th data point.

- ​ represents the predicted value of the target variable for the \( i \)th data point.

The process involves calculating the squared difference between each predicted value and its corresponding actual value, summing up these squared differences across all data points, and finally averaging the result. The resulting MSE value provides an overall measure of the model's ability to accurately predict the target variable across the dataset.

Interpreting MSE involves understanding its properties and implications. A lower MSE indicates better predictive performance, as it suggests that the model's predictions are closer to the actual values. Conversely, a higher MSE signifies poorer performance, indicating larger discrepancies between predicted and actual values.

MSE is particularly useful for comparing the performance of different models or evaluating the impact of various hyperparameters and architectural choices. By calculating MSE for multiple models or variations of a single model, researchers can identify which configurations yield the most accurate predictions.

However, it's essential to consider potential limitations when interpreting MSE. For instance, MSE is sensitive to outliers, as the squared differences heavily penalize large errors. In scenarios where outliers are prevalent or influential, alternative metrics such as Mean Absolute Error (MAE) may provide a more robust evaluation.

In summary, MSE serves as a valuable tool for assessing the accuracy of predictive models, including those used in time series forecasting. Its straightforward calculation and intuitive interpretation make it a popular choice for researchers and practitioners seeking to quantify the performance of their models accurately.

IV.EVALUATION

IV.I Evaluation in Long short-term memory(LSTM)

In the experiment conducted, several key values and components are utilized in training an LSTM model for time series forecasting. One crucial aspect is the sequence length, which determines the length of historical data sequences fed into the model for prediction. By defining an appropriate sequence length, the model can capture temporal patterns and dependencies in the data effectively. In this case, a sequence length of 10 is chosen, but this value can be adjusted based on the specific characteristics of the dataset and the forecasting task at hand.

Additionally, the experiment splits the dataset into training and testing sets, ensuring that the model's performance can be evaluated on unseen data. This practice helps assess the generalization ability of the model and guards against overfitting, where the model memorizes the training data but fails to perform well on new data. By separating the data into distinct subsets, researchers can accurately gauge the model's predictive accuracy and identify any potential issues.

Another critical aspect of the experiment is the creation of a reusable class for defining the LSTM model architecture. By encapsulating the model within a class, it becomes modular and can be easily instantiated and reused for different parts of the analysis. In this case, the same LSTM class is employed for both the American and British parts of the analysis, highlighting the efficiency and flexibility of object-oriented programming. This approach not only streamlines the code but also promotes code reusability and maintainability, as modifications or updates to the model can be applied uniformly across multiple instances.

Furthermore, the experiment demonstrates the importance of data preprocessing and conversion to PyTorch tensors before training the model. PyTorch tensors provide a flexible and efficient way to perform numerical computations, particularly when leveraging GPU acceleration for deep learning tasks. By converting the input and output data into tensors, the experiment ensures compatibility with PyTorch's computational graph framework, facilitating efficient gradient computation during the training process.

IV.II Evaluation in Gated Recurrent Unit (GRU)

The experiment conducted in this research entails the implementation of a Gated Recurrent Unit (GRU) model, tailored to forecast the value of merchandise trade between Ireland and various countries, with a specific focus on the USA and Great Britain. GRU models represent a type of recurrent neural network (RNN) renowned for their adeptness in handling sequential data, a pertinent trait considering the temporal nature of trade datasets. Within the GRU class definition, three primary components are delineated: the GRU layer, the linear layer, and the hidden cell. These elements collectively facilitate the model's ability to process sequential inputs, map hidden states to output dimensions, and initialize the hidden state, respectively.

In the context of the research, the importance of this class lies in its capacity to encapsulate the architecture and functionality of the GRU model, which forms the cornerstone of the predictive framework. Through the `forward` method, the model undertakes the forward pass, wherein input sequences are processed by the GRU layer to generate predictions via the linear layer. This method culminates in the derivation of the final prediction, corresponding to the last element of the output sequence, thus encapsulating the predictive capacity of the model.

To facilitate model training, separate instances of the GRU class are instantiated for the American and British trade segments. Distinct optimizers are also designated to update the parameters of each model during the training phase. The training process itself, orchestrated by the `train\_model\_gru` function, unfolds iteratively over the training data across specified epochs. Within each epoch, the function iterates through sequence-label pairs, computes predictions, evaluates loss using the mean squared error (MSE) metric, and updates model parameters via backpropagation.

In the subsequent evaluation phase of the research, the effectiveness of the GRU model is scrutinized. Central to this assessment is the quantification of the model's predictive accuracy through the MSE metric, which gauges the disparity between predicted and actual trade values. By comparing the MSE obtained from the GRU model trained on American and British trade data, researchers gain insights into the model's efficacy in discerning and forecasting trade dynamics between diverse countries. This evaluative framework furnishes valuable insights into the utility of GRU models in forecasting international trade patterns, thereby enriching our understanding of economic dynamics on a global scale.

V. RESULTS

V.I LSTM part:

The LSTM model is trained separately for both the American and British parts of the analysis. After training, the loss values for each epoch are logged to monitor the model's convergence and performance.

For the American part, the output of the model training function reveals a gradual decrease in the loss values over epochs, indicating improved accuracy in predicting the merchandise trade values between Ireland and the United States. Specifically, the loss values decrease from 0.00338980 at epoch 0 to 0.00025278 at epoch 125. This decline in loss signifies that the model successfully learns the temporal patterns in the American trade data and effectively predicts future values.

Similarly, for the British part, the loss values exhibit a decreasing trend over epochs, demonstrating the model's ability to capture the underlying patterns in the merchandise trade between Ireland and Great Britain. Starting from 0.00394098 at epoch 0, the loss decreases to 0.00283205 at epoch 125. This decline in loss indicates that the model learns to make increasingly accurate predictions for the British trade relationship as training progresses.

By monitoring the loss values during training, researchers gain valuable insights into the model's convergence and predictive performance. The observed decrease in loss signifies that the LSTM model effectively learns from the input data and makes accurate predictions for both the American and British trade relationships. These findings highlight the efficacy of LSTM in capturing the complex temporal dynamics inherent in international trade data and underscore its utility in forecasting future trade values.

After utilizing LSTM to forecast the values of the test data, we generated the following graph to visualize the predicted outcomes:

V.II GRU part:

After training the GRU model on the dataset pertaining to the American trade segment, the results demonstrate a fluctuating pattern in the loss function across epochs. Initially, during the initial epochs, the loss exhibits a slight increase from 0.00053332 to 0.00059335, signifying a minor divergence between predicted and actual values. However, as the training progresses, the loss fluctuates, with occasional spikes and dips observed. Notably, around the 50th epoch, a significant reduction in loss is witnessed, dropping to 0.00033663. This decline suggests an improvement in model performance, characterized by enhanced predictive accuracy. Despite fluctuations in subsequent epochs, including a noticeable increase at the 100th epoch, the overall trend showcases a reasonable level of convergence, culminating in a relatively low loss value of 0.00009290 by the 125th epoch. This outcome implies that the GRU model has effectively learned the underlying patterns within the American trade data, yielding promising predictive capabilities.

In contrast, the training results for the British trade segment reveal a somewhat different trajectory in model performance. The loss function exhibits a comparable initial trend, with a gradual decline observed from 0.00271742 to 0.00215432 across the first 50 epochs. This reduction signifies a progressive improvement in predictive accuracy, albeit at a slower pace compared to the American trade segment. Notably, a significant drop in loss occurs around the 100th epoch, plummeting to 0.00050506. This substantial decrease suggests a notable enhancement in the model's ability to capture and predict trade dynamics specific to the British market. Furthermore, the subsequent epochs demonstrate a consistent stabilization in loss, with minimal fluctuations observed. By the 125th epoch, the loss value further diminishes to 0.00045351, underscoring the model's adeptness in accurately forecasting trade values within the British context. Overall, the results highlight the efficacy of the GRU model in capturing distinct trade patterns across different geographical regions, thus underscoring its utility in international trade forecasting endeavors.

After utilizing GRU to forecast the values of the test data, we generated the following graph to visualize the predicted outcomes:

V.III MSE

The comparison of mean squared error (MSE) between LSTM and GRU models provides valuable insights into their respective performance in forecasting merchandise trade values for both the USA and Great Britain. In the case of the USA, the LSTM model yields an MSE of 0.02861, while the GRU model achieves a notably lower MSE of 0.02000. This discrepancy suggests that the GRU model outperforms its LSTM counterpart in predicting trade values for the American market. The lower MSE attained by the GRU model indicates a higher degree of accuracy in capturing the underlying trade dynamics and trends specific to the USA, showcasing its effectiveness in this forecasting task.

Conversely, when examining the British trade segment, the LSTM model exhibits an MSE of 0.05120, whereas the GRU model records a comparatively higher MSE of 0.09169. Unlike the results observed for the American part, the GRU model's performance lags behind that of the LSTM model in forecasting trade values for Great Britain. The higher MSE obtained by the GRU model suggests a diminished level of predictive accuracy relative to the LSTM model within the British context. This outcome underscores the nuanced nature of model performance across different geographical regions, highlighting the importance of considering regional-specific factors and dynamics in trade forecasting endeavors.

In summation, the comparison of MSE values between LSTM and GRU models underscores the variability in their performance across distinct trade segments. While the GRU model demonstrates superior predictive capabilities for the USA, evidenced by its lower MSE compared to LSTM, it exhibits diminished performance relative to LSTM in forecasting trade values for Great Britain, as indicated by its higher MSE. These findings underscore the significance of selecting an appropriate model architecture tailored to the specific characteristics and dynamics of the trade data under consideration, thereby enhancing the efficacy of forecasting endeavors in international trade analysis.

TABLE I.

|  |  |  |
| --- | --- | --- |
|  | USA | Great Britain |
| MSE in LSTM | 0.02861 | 0.05120 |
| MSE in GRU | 0.02000 | 0.09169 |

VI. CONCLUSION

In conclusion, the analysis of mean squared error (MSE) metrics based on LSTM and GRU models provides valuable insights into their effectiveness in forecasting merchandise trade values between Ireland and different countries, particularly the USA and Great Britain. Through this comparative analysis, several key findings have emerged, shedding light on the performance of each model architecture within distinct trade segments.

Firstly, the LSTM model demonstrates superior predictive capabilities for the British trade segment, as evidenced by its lower MSE compared to the GRU model. This finding suggests that LSTM's architectural nuances may be better suited to capture the complex temporal dependencies inherent in the trade dynamics specific to Great Britain. Conversely, the GRU model excels in forecasting trade values for the USA, as indicated by its lower MSE relative to LSTM. This outcome underscores the efficacy of GRU's gating mechanisms in discerning and predicting the underlying patterns within the American trade data.

Furtheremore, the contrasting performance of LSTM and GRU models underscores the importance of selecting an appropriate model architecture tailored to the unique characteristics and dynamics of each trade segment. While LSTM may excel in capturing certain temporal dependencies prevalent in one region, GRU may prove more adept at discerning patterns specific to another. Thus, the choice between LSTM and GRU models should be informed by a nuanced understanding of the underlying trade dynamics and contextual factors inherent in each geographical region.

VII. DISCUSSION

##### References

[1] M. C. Jensen, "Global Trade and the Role of State-Owned Enterprises," Harvard University Press, 2018.

[2] J. Frankel, "The Globalization of International Trade: Growth, Poverty and Inequality," Springer, 2019.

[3] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, "Time Series Analysis: Forecasting and Control," Wiley, 2015.

[4] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

[5] Y. Bengio, "Deep Learning," MIT Press, 2016.

[6] F. J. Oles, "Introduction to Econometrics," Addison-Wesley, 2000.

[7] M. Zaharia et al., "Apache Spark: A Unified Engine for Big Data Processing," Communications of the ACM, vol. 59, no. 11, pp. 56–65, 2016.

[8] T. White, Hadoop: The Definitive Guide, O'Reilly Media, 2015.

[9] V. Gulzar, "Big Data Analytics Using Apache Spark and Hadoop," ProQuest Dissertations Publishing, 2017.

[10] T. Nadeau and M. Reddy, Hadoop: Beginner's Guide, Packt Publishing Ltd, 2013.

[11] S. Karau et al., High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark, O'Reilly Media, 2017.