ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BEYOND THE NOISE: UNVEILING MEANINGFUL PATTERNS IN MOBILE PAYMENT DATA USING FUZZY LOGIC AND SVM

GAITH M. ABASHES [1], BASIMA ELSHQEIRAT[2] , AHMAD A. ABU-SHAREHA[3], MOHAMMAD ALSHRAIDEH[4]

[1,2]Computer Science Department, The University of Jordan, Amman, Jordan
[3]Department of Data Science and Artificial Intelligence, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan.
[4]Artificial Intelligence Department, The University of Jordan, Amman, Jordan
Email: [1]gaithh1989@yahoo.com , [2]b.shoqurat@ju.edu.jo , [3]a.abushareha@ammanu.edu.jo , ,
[4]mshridah@ju.edu.jo

ABSTRACT

Big Data refers to vast and complex datasets that require processing and analysis to uncover valuable insights beneficial to businesses and organizations. This has become a fundamental approach for acquiring, processing, and analyzing large volumes of data to extract useful information. Big Data encounters challenges such as Volume, Velocity, Variety, Variability, and Value. Preprocessing and analysis are crucial for obtaining quality information that supports accurate decision-making. Many organizations now handle extensive data, often called "big data," due to its size, speed, and diverse formats, revolutionizing decision support and data management. The primary challenge of big data is to extract values for decision-making, prediction, and service improvement.

The proposed framework focuses on preserving financial data quality by clustering with fuzzy Logic and Support Vector Machines (SVM). It comprises four layers: data collection in the first layer, preprocessing data in the second layer, which involves cleaning and mapping semi-structured data to establish relationships; the third layer applies a fuzzy controller and classification to generate rules; the fourth and final layers; and data reduction and classification employing SVM clustering to create distinct clusters for meaningful and predictive outcomes. This study addressed two key challenges:

1. Extracting meaningful information from big data: This study aimed to extract valuable insights from the vast data generated in the mobile payment sector.
2. Utilizing big data techniques with SVM clustering and leveraging fuzzy Logic, this study employs Support Vector Machine (SVM) clustering to identify patterns and relationships within the data. Additionally, fuzzy Logic was incorporated to extract the rules and connections among the attributes.

We experimentally validate the proposed approach using financial company data from the mobile payment industry.

Keywords: *Big Data, Decision Support, Prediction, Fuzzy Logic, Support Vector Machines (SVM).*

1. INTRODUCTION

The domain of big data encompasses a wide range of data types, including structured, unstructured, and semi-structured data sourced from diverse fields, such as industry, business, social networks, the Internet, healthcare, finance, economics, and transportation. In the current era of big data, the necessity for adaptable tools and techniques is paramount, as the sheer volume and complexity of big data present challenges for traditional data-processing tools. Big data refers to the accumulation of extensive and intricate datasets that surpass the capabilities of conventional data-processing tools and relational database management systems' capabilities in processing, managing, and capturing data within reasonable timeframes. The classification of datasets as big data depends on the organization's size. The term' big data' has permeated various industries and academia, and its challenges include data capture, storage, search, sharing, transfer, analysis, and visualization.

Sam Madden [14] from the Massachusetts Institute of Technology (MIT) characterizes big data as being "too big, too fast, or too hard" for existing tools to process. This refers to data that may reach the petabyte scale, originate from diverse sources, proliferate, and present challenges not neatly aligned with existing processing tools.

Contemporary organizations amass, store, and analyze massive volumes of data, often termed "big data" due to its vastness, speed, and varied formats [13]. This influx of big data heralds a new era in decision-support data management. Realizing the potential value of these data, organizations invest in technologies, personnel, and processes to harness these opportunities. The crux of deriving value from big data lies in analytics: merely collecting and storing it does not create significant value. Instead, the data must be analyzed, yielding results that decision-makers and organizational processes can leverage to generate value.

To address the challenges posed by big data, several MapReduce frameworks have been developed, including Apache Hadoop, Skynet, Sailfish, and file maps [5]. These frameworks efficiently handle large and immutable datasets (e.g., logs or large binary objects) and incrementally collected data (e.g., web crawls, user comments on social networks, GPS data, and sensor events). They find applications in diverse use cases, such as log file analysis, scientific simulations, and financial predictions.

Big data's primary challenge is extracting value to facilitate decision-making, prediction, and service improvement. Traditional data-mining techniques struggle with big data, necessitating the application of artificial intelligence techniques. Big data are often characterized by the "5Vs": velocity, volume, value, variety, and veracity [7].

This research aims to introduce a robust framework strategically crafted to aid companies in acquiring high-quality information, fostering precise decision-making, and yielding valuable insights that prove advantageous for businesses and their clientele. This study is meticulously geared towards achieving specific objectives, notably the development of an innovative, extensive data framework explicitly designed for the scrutiny and interpretation of financial data. The primary focus is on extracting pertinent information that serves as a foundation for informed decision-making processes within the intricate landscape of financial institutions.

This study makes notable contributions by incorporating a Support Vector Machine (SVM) and fuzzy logic algorithms into data testing and analysis. SVM, known for its advantages in classification and regression tasks, enhances the framework's capability to discern patterns and relationships within financial datasets. Concurrently, integrating fuzzy logic algorithms provides a nuanced approach to handling uncertainty and imprecision in data, thereby enriching the analysis process.

Moreover, the study undertakes a practical dimension through the execution of experiments specifically designed to showcase the effectiveness of the proposed framework. By applying the developed framework to real-world scenarios using financial company data in the dynamic mobile payment sector, this research endeavors to demonstrate its practical utility and efficacy. The results of these experiments aim to validate the framework's ability to derive meaningful insights, support accurate decision-making, and contribute positively to the overall efficiency of financial processes.

Essentially, this study represents a substantial effort to advance methodologies for data analysis in the financial domain. This underscores the integration of cutting-edge techniques to improve decision-making processes and cultivate a more informed and efficient financial landscape.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive background on Big Data, including its characteristics, types, architecture, frameworks, and data testing and analysis tools. Section 3 delves into the proposed methodology, presenting the "framework" applied to test and analyze data. This Section also offers a detailed account of the algorithms used and the data involved in the study. Section 4 presents the experimental results, evaluation criteria, and assessment of the findings. Section 5 provides an overview of the research and explores avenues for future research.

2. BACKGROUND AND LITERATURE REVIEW

The term' big data' has gained prominence in various industries and academia, but its definition remains ambiguous and evolving. This ambiguity poses challenges for its consistent use and development. Establishing a standard definition is crucial for ensuring orderly growth and reducing confusion. Although there are multiple definitions, the core idea is that big data refers to extensive and complex datasets that traditional data management tools cannot efficiently store or process. These datasets require processing and analysis to extract valuable insights beneficial to businesses and organizations. A clear and widely accepted definition is vital to the scientific progress of this concept [2].

Classification is fundamental to understanding big data and is typically categorized into structured, unstructured, and semi-structured. Consensus in the field revolves around the four Vs defining big data: volume, variety, velocity, and veracity. Additionally, data can originate from diverse sources and fields, including financial companies (e.g., the New York Stock Exchange, Jordanian mobile payment companies), social media platforms (e.g., Facebook, Twitter, Instagram), and even single jet engines on airplanes. Big data offers advantages such as improved customer service, churn prediction, healthcare, location-based services, crime prevention, operational efficiency, and informed decision-making.

Recent publications have focused on addressing significant data challenges, particularly emphasizing the four V's. Additionally, there is a growing focus on improving the extraction of value from big data through timely responses. Various approaches have been proposed to address big data classification problems, highlighting the importance of practical data analysis and utilization in different domains [3, 7], as shown in Table 1. In a study conducted by [12], a framework was presented to enhance the reduction of large data dimensions. This framework focuses on selecting more essential features to improve the accuracy of the decision-tree-based rating performance. The results showed an increase in accuracy from 84.3% to 86.4% after implementing this feature selection approach. The data processing step addressed two significant issues: handling heterogeneous data through peer-to-peer switching and dealing with incomplete data using fixed number assignments. In the mapping step, feature selection was accomplished using fuzzy rough sets. In the reduce step, fuzzy applications facilitate clustering to identify similar features and assign them to the same key.

Table 1: Comparison Among Similar Approaches

Problem	Approach	Case study	Tools
Big data dimensionality	MapReduce parallel processing and fuzzy rough	Diabetes dataset electroencephalography	WEKA
Big data classifying problems.	MapReduce approach with dynamic fuzzy inference	Six UCI datasets	JAVA
Extraction of information	The Chi et al.'s algorithm for classification	Six problems from the UCI dataset repository	JAVA

We used MapReduce with dynamic fuzzy interpolation/interpolation in large data applications, which allowed for the integration of reasoning and completion techniques, resulting in final outputs. Empirical research involving six different big data problems, as conducted by [11], demonstrated varying average accuracy in performance between the two methods.

This study by [22] employs exploratory data analysis on balance sheets, income statements, and cash flow data, using parameters like Debt-to-Equity Ratio, Current Ratio, Return on Capital Employed, Net Profit Margin, and Inventory Turnover Ratio for investment decisions. Predictive analytics uses four machine learning models: linear regression, K-nearest neighbor, support vector regression, and decision tree. Results suggest that the decision tree is the most valuable for performance analytics, with the optimal hyperparameter (maximum depth) determined as nine through a grid search.

The article in [24] explores the contemporary relationship between finance and big data, focusing on its impact on financial markets, institutions, internet finance, credit services, fraud detection, risk analysis, and application management. A literature review highlights the intricate connections between big data and various financial components, providing a foundation for future research directions in this evolving field.

The authors in [23] introduced a new extensive data processing framework to address the challenges of efficiently handling this data. The framework provides a dual-mode solution for processing data in historical and real-time scenarios, featuring optimized functional modules. Comparative tests show that the proposed framework outperforms existing alternatives, demonstrating its effectiveness in meeting the unique data processing needs of the Online Roadshow.

The inherent uncertainty and noise in the available data challenge big data classification problems, particularly in information extraction. To evaluate the performance of the ChiFRBCS data algorithm in large data scenarios, del Río et al. [16] assessed the rating based on the achieved resolution and the operating time required for the models. This analysis aimed to gauge the quality of the algorithm's performance in handling extensive data.

Big data is characterized by its vastness and complexity, necessitating new technologies and structures for effective capture and analysis. Traditional techniques struggle to handle the sheer volume and diverse characteristics of big data, which include size, speed, variety, variability, and value. This paper underscores the importance of big data technology in the modern world and highlights projects transforming science and society.

Based on [17], data growth is rapid, reaching exabytes, and poses challenges due to the exponential increase in volume compared to available computing resources. While the term "big data" primarily refers to volume, it encompasses several properties:

1. Variety: Big data consists of diverse data types, including structured, semi-structured, and unstructured data from various sources such as web pages, log files, social media, emails, documents, and sensor data. Managing and analyzing this heterogeneous data is a challenge.
2. Volume: Big data represents an enormous volume of data, measured in bytes and potentially reaching zettabytes soon. Social networking sites alone generate terabytes of data daily, straining traditional systems.
3. Velocity: Velocity refers to the speed at which data arrives from different sources and how it flows continuously. Sensor data, for instance, streams into databases without pause, requiring analytics capable of handling this constant flow.
4. Variability: Variability encompasses the inconsistencies in data flow, often exacerbated by events like social media spikes. Maintaining data consistency becomes challenging in such scenarios.
5. Value: Extracting value from big data involves running queries to derive essential insights and trends, aiding in strategic decision-making.

Significant data types differ from those stored in traditional warehouses. Traditional data must be well-documented, trusted, and stored in a structured format aligned with the warehouse infrastructure. In contrast, big data encompasses conventional and unconventional data types, making it accessible for analysis and enabling improved business and decision strategies.

Big data is commonly classified into three main types:

1. Structured Data: Data with a fixed format that is quickly processed, stored, and retrieved. Structured data resembles organized database tables.
2. Unstructured Data: Data lacks a specific structure, making it challenging and time-consuming to process. Examples include email content.
3. Semi-Structured Data: Semi-structured data refers to information that does not fit neatly into the traditional structure of a relational database, yet it has some level of organizational structure. Unlike structured data with a rigid schema or unstructured data lacking a predefined schema, semi-structured data falls in between. An example of semi-structured data is JSON (JavaScript Object Notation) format.

The classification of data types serves as a foundational aspect of studying big data, with each type presenting its unique challenges and opportunities, as shown in Table 2 based on [18].

Source of Big Data: Big data is distinct from data stored in traditional warehouses, requiring different approaches to management. Various new data sources have contributed to the significant growth of big data technology investments. These sources include industries digitizing their content, leading to rapid data growth. Critical sources of big data include [4]:

- **Log Storage in IT Industries:** IT industries store large amounts of data as records to address occasional issues. Due to its size and raw nature, data is often stored for short durations. Big data analytics can help analyze this data thoroughly and extend its storage life.
- **Sensor Data:** Managing vast amounts of sensor data presents challenges, as many industries only use a fraction of this data for analysis due to storage and analysis limitations.
- **Risk Analysis:** Financial institutions use data modeling to calculate risk, necessitating incorporating significant, underutilized data into risk pattern identification.

- **Social media:** Big data widely monitors customer sentiments and feedback, informing business decisions.
- **Healthcare:** The healthcare industry is transitioning to electronic medical records and images for public health monitoring and epidemiological research.

Big Data Challenges and Issues: Big data analysis involves distinct phases, each presenting unique challenges. Some significant challenges and issues include [9]:

- **Privacy and Security:** Protecting sensitive big data presents significant conceptual, technical, and legal challenges.
- **Data Access and Sharing:** Timely, accurate, and complete data access is crucial for informed decision-making. Data sharing can be complicated but is essential for improved decision-making and productivity.
- **Storage and Processing:** Storing and processing vast data volumes is challenging and requires efficient data transfer, storage, and processing methods.
- **Analytical Challenges:** Analyzing unstructured and semi-structured big data requires advanced skills and methods, with analysis often dependent on specific results and decision-making goals.
- **Skill Requirements:** Organizations require diverse skill sets, including technical, research, analytical, interpretive, and creative skills, to harness the potential of big data.
- **Technical Challenges:** Fault tolerance, scalability, data quality, and heterogeneity pose significant problems in testing and analyzing big data.

What is Big Data Architecture? Significant data architecture is designed to manage large and complex data sets' capture, storage, and analysis. It encompasses various components, including as shown in Figure 1 [19]:

- **Data Sources:** The starting point of a significant data pipeline, including all data sources.
- **Data Storage:** Utilizes distributed file stores and data lakes for batch-based operations.
- **Batch Processing:** Segregates data into chunks, processes, filters, and aggregates.
- **Real-Time-Based Message Ingestion:** Handles real-time data flows and message ingestion, often using systems like Apache Kafka or Apache Flume.
- **Stream Processing:** Manages streaming data for windowed or sequential data flow analysis.
- **Analytics-Based Datastore:** Used for analytical purposes, querying, and analyzing processed data.
- **Reporting and Analysis:** Provides insights and reports using business intelligence tools.
- **Orchestration:** Manages repetitive data operations and workflows.

Big Data Testing and Analysis: Big data analysis relies on recent technological advancements, capturing and analyzing high-speed data. Data sources extend beyond traditional databases to include unstructured data from emails, mobile outputs, and sensor-generated data. Successful considerable data testing requires a deep understanding of data storage, business intelligence, and the technologies used in big data frameworks. Building skilled test teams with coding, white box testing, and data screening capabilities is crucial. Testing big data is essential for identifying quality issues and ensuring optimal data utilization [10].
Big data analytics offers numerous advantages for organizations, including improved sales, efficiency, customer service, and risk management, ultimately enhancing profitability and agility.

3. PROPOSED FRAMEWORK

This Section introduces the proposed Big Data framework for data analysis and testing. It explains the framework's significance in the research context and details the data types and structures employed for analysis.
The proposed framework consists of five primary layers:

1. Data Collection Layer: This layer focuses on gathering the required data for analysis.
2. Data Preprocessing Layer: Here, semi-structured data is converted into a

structured format and prepared for input into subsequent layers.

3. Rule Extraction Layer: This layer extracts rules and relationships between data tables.
4. Clustering Layer: This layer performs data grouping (clustering) using the SVM algorithm.
5. Data Analysis and Testing Layer: The final layer analyzes and tests data to extract meaningful insights.

Each of these layers addresses specific characteristics of big data, such as volume, variety, velocity, and value. Figure 2 illustrates the entire process of the proposed framework.

This Section provides detailed explanations of each layer within the framework, offering a comprehensive understanding of how the framework operates.

3.1 Data Capture and Data Collection:

Big data, a complex interplay of dataset characteristics, analysis, system performance, and cost-effective business considerations, exceeds typical database tool capacities. It involves diverse data collection sources, from laboratories to online interactions. Even traditional industries benefit by integrating external data. Data capture includes encryption for numerical value assignment to responses, whether automated or manual.

This study will leverage customer data from a financial firm specializing in mobile payment wallets. The datasets encompass customer transaction details (financial data), customer information (non-financial data), and mobile application usage patterns. Spanning a specified period, these datasets aim to yield valuable insights into customer behavior, financial transactions, and personal information.

3.2 Data Preprocessing:

After collecting data, diverse data sources with varying characteristics are expected to be encountered. The immediate goal is to standardize these sources for the continued development of our data product. However, whether to homogenize the data depends on its nature. We need to consider if it is practical to do so.

Data cleansing involves identifying and rectifying errors, such as inaccuracies, incompleteness, or unreasonable entries. Various methods detect outliers, such as statistical analysis, pattern recognition, and correlation. Additionally, error detection techniques, data formatting conversions, safety restrictions enforcement, deduplication, and other statistical methods contribute to data auditing [6].

The cleansing process follows a logical sequence of steps to detect and rectify errors and inconsistencies to enhance data quality. Data quality issues often arise within individual datasets due to spelling errors, missing information, or invalid entries. When consolidating multiple data sources into data warehouses or global information systems, the need for data cleaning becomes even more pronounced. This is because these resources often contain redundant data in various formats. Therefore, consolidating different data representations and eliminating duplicate information is essential for providing accurate and consistent data access [15].

Data cleaning is crucial for maintaining data consistency and accuracy, although it involves computational overhead and data density. Due to privacy concerns, sensitive data like names, national IDs, phone numbers, etc., are not required for this thesis. Customer information includes attributes like customer status, Gender, nationality, ID type, date of birth, City, and mobile number (MNO operator) along with customer profiles. Customer transactions include transaction date, direction, status, sender info, receiver info, and endpoint operation.

Preprocessing techniques, such as removing duplicate rows and spaces between data content, are applied to ensure the efficiency and quality of subsequent analysis processes.

This thesis employs preprocessing techniques to streamline data analysis and algorithmic adaptation. After applying various preprocessing steps, the resulting data is considered high-quality and reliable, suitable for data analysis, aggregation algorithms, and other preprocessing technologies.

Key actions in data preprocessing include:

- Ensuring compliance with privacy policies by excluding sensitive data.
- defining variables for ease of handling and analysis.
- Removing spaces between data content to avoid negative impacts on results.
- Eliminating duplicated and unnecessary data.
- Deleting records with null or zero attribute values.

Moreover, the 'Preprocess' package within the R framework reduces data volume and ensures clean input for subsequent layers [1]. This stage focuses on attributes and content to ensure clean data for downstream processing.

3.3 Extraction Rule and Relation:

The preprocessing results will serve as input for the extraction acquisition layer. Data acquisition encompasses collecting, filtering, and cleaning data before storing it in a data repository or other storage solution for analysis. Managing extensive data presents infrastructure challenges, necessitating low and predictable latency, scalability for high transaction volumes, and support for dynamic data structures [25].

A vital data feature is the customer's outcome—whether they are active or have terminated their relationship with the company. This attribute facilitates comparative analysis across different applications and algorithms. Big data analysis tools like Spark, R, Hadoop, Storm, HPCC, and Cassandra are available. For this research, data, and transactions from a local financial company specializing in mobile payments over a specific period, such as two years, are utilized [26].

Two primary tools are used for data analysis: Spark with MapReduce algorithm and R language with Fuzzy Logic and Association Rule algorithms. These tools extract rules and relationships from customer transactions, information, and messages based on attribute strength and usability. Fuzzy Logic and Association Rule algorithms predict improvements for the company's services based on attribute relationships and enforce recovery rules for split data sets. Rules and relationships may involve one or more attributes from the same or different groups, depending on their correlation strength. Data analysis considers standard thresholds of significance to assess the impact of values on each group [26].

Table 3 illustrates the main affected rules and relationships to be used in the data analysis process.

Table 2: Main Effected Rules

Rule ID	Rule description
1	Customer Status Based on Gender
2	City based on Gender
3	MNO (Operator) based on Gender
4	ID Type based on Gender
5	Relation between City, Gender, and ID Type "Customer for each city with gender and ID Type"
6	Relation between City, Gender, and MNO "Customer for each city with gender and MNO (Operator)"
7	Relation between City, Gender, and Customer status "Status of Customer based on City and Gender"
8	Number of transaction peer city based on status
9	Number of transaction peer city based on direction
10	Number of endpoint operations based on status

3.4 Data Grouping:

Big data represents unstructured information [20]:

3.4.1 Clustering Algorithm:

Cluster algorithms, a subset of machine learning algorithms, are employed to partition a dataset into distinct groups based on specific criteria and business needs. Integral to data science and artificial intelligence (AI), these algorithms have diverse applications. They fall into two main types: rigid aggregation and flexible grouping. Standard clustering methods, determined by the calculation process, encompass K-Means, connection models, Centroid models, distribution models, density models, and hierarchical grouping. These algorithms find utility in various domains, including image segmentation, market segmentation, and social network analysis, as shown in Figure 3 [27].

3.4.2 Classification Algorithm:

Classification is a data analysis method that categorizes data into predefined classes, assigning labels based on the analysis of a training dataset. The main goal is to organize new data points accurately. The process involves a learning phase, where the model is trained, and an assessment phase, where it predicts outputs for new inputs. In healthcare, for instance, classification is used to diagnose patients by analyzing data such as names, addresses, ages, and health histories. This involves creating a mathematical function denoted as "F" with inputs ("x") and outputs ("y"). Classification algorithms have widespread applications, including spam detection, bank loan approval prediction, speech recognition, and sentiment analysis, as shown in Figure 4[21].

In this study, we utilize the Support Vector Machine (SVM) algorithm in conjunction with the Correlation technique for data classification and grouping, employing Spark and R tools. Data classification is executed within this layer using the 'e1071' Package in the R framework, accessible at (https://cran.r-project.org/web/packages/e1071/index.html).
Support Vector Machine (SVM) is a supervised learning technique for solving two-class classification problems. Paired with the Correlation technique, it identifies relationships between variables within a dataset [1].
SVM, or Support Vector Machine, is a versatile machine learning algorithm for data classification and regression analysis. Initially tailored for two-class classification problems, SVM creates a data map with maximal margins between classes. Its applications span text categorization, image classification, handwriting recognition, and scientific research, effectively handling linear and non-linear problems [1].
Correlation, on the other hand, is a technique for discerning relationships between variables in a dataset. Quantified by a mathematical value from -1 to +1, correlation signifies the strength and nature of the relationship without implying causation. For example, the other may increase or decrease as one variable increases. It is essential to note that correlation does not reveal the root cause of a relationship. With a rich history, the correlation method is crucial for the accurate and efficient analysis of large datasets, playing a pivotal role in the future of scientific research [8].

4. EXPERIMENTAL RESULTS AND EVALUATION

This Section delves into the "Approach" framework for extracting meaningful data using R and Spark packages. It covers the experimental setup and evaluation parameters before presenting the results and conducting an analysis.

4.1 Simulation Environment

The experiments were conducted independently on the same computer system configurations, using Spark and R tools to test and analyze the data.

4.2 Data Capture and Collection

In this research, we rely on customer data from a financial company specializing in mobile payment wallets. Our focus is on various aspects of customer data, including customer transactions (financial data), customer information (non-financial data), and customer behavior within the application (messages). This data is sourced from users based on their transactions and usage patterns. The dataset is provided by a local financial company specializing in mobile payment methods and spans a specific timeframe, typically two years. This data set contains extensive information about clients, their financial transactions, and their interactions with the application.
In Spark, we imported the dataset using the "import data" function, which retrieves data from a file and enables a comprehensive view and understanding of the entire dataset. One advantage of R packages is their capability to handle both in-memory and out-of-memory storage, making them suitable for dealing with large volumes of data.
The "Customer Information" data set comprises 108,980 records, with all associated properties detailed in Table 4.

Table 3: Description Of Customer Information

Attribute Name	Data Type	Description
Customer Status	Character	Status on the system
First Name	Character	The user's first name
Last name	Character	The user's last name
Gender	Character	Determine the Gender of the customer
IS Active	Character	Is the customer active or not
Nationality ID	Integer	Determine the customer's nationality
ID Type	Character	the document in which the customer is registered

		Ex: National ID, Passport
ID Number	Integer	The document number with which the customer is registered
Date of Birth	Date time	The customer's birthday
Client Type	Character	Determine the type of customer Ex, Person, Company
Client Reference	Integer	Where the customer was registered
IS Registered	Character	If the customer registered or not
City	Character	City for customer
Mobile Number	Integer	Mobile number for customer
MNO	Character	Telephone network where the customer is registered EX: Zain, Orange, Umniah
Customer Profile	Character	Name of profile related to customer

- Transaction information consists of **564912** records with all properties, as shown in Table 5.

Table 5: Description of Transaction information

Attribute Name	**Data Type**	**Description**
Reference	Integer	Transaction number
Transaction Date	Date time	Transaction Date
Direction	Character	The direction of the transaction Ex: Outward, Onus, Inward
Status	Character	The status of the transaction Ex: Success, Reject,
Reason Description	Character	Reason for transaction
Receiver Info	Integer and Character	Receiver info for transaction
Sender Info	Integer and Character	Sender Info for transaction
Source Name	Character	Name of customer who is doing the transaction
Destination Name	Character	Destination name for transaction
Original Amount	Integer	Original amount for a transaction without fees
Total Amount	Integer	Total amount for transaction with fees
External Fees	Integer	The fees for a transaction
End Point Operation	Character	Determine the

		operation for transactions, Ex Money Transfer, Agent cash-in
Is Reversed	Character	If the transaction Reversed or not, Yes or No
City	Character	The City of the customer does the transaction

- The message consists of **2632555** records, as shown in the table below, with all properties in Table 6.

Table 6 Description of Messages information

Attribute Name	**Data Type**	**Description**
Sender	Integer	Which customers do the transaction
Message Type	Character	Type of message Ex: P2P, Cash-in, Change Alias
Operation	Character	Type of operation Ex: Money Transfer, Alias Change
Processing Status	Character	The status of the operation
Processing Stamp	Date time	The time of operation

4.3 Data Preprocessing

After comprehensively examining the data, the data preparation phase addresses all potential influencers on the results, including null values, outliers, and missing data. Preprocessing aims to manage the inherent uncertainty introduced by the diversity and integrity of big data while meeting volume and velocity requirements. This stage involves two primary steps: removing missing values and data conversion.

In Spark, preprocessing utilizes the MapReduce algorithm and built-in functions, while in R, the Preprocess package is employed to ready the data for analysis. Various preprocessing techniques are applied to streamline data analysis, aligning it with the algorithm. The resultant data, considered high-quality and reliable, is prepared for subsequent analysis and clustering algorithms.

Critical steps in the preprocessing layer involve removing missing values in selected attributes, data conversion, and attribute selection. Attribute selection is pivotal, filtering data based on relevance and content. This process identifies and retains essential characteristics for the specific analysis area while eliminating irrelevant or redundant features.

Tables 7, 8, and 9 below showcase the data and attributes after applying preprocessing techniques.

- The "Customer Information" data set comprises 108,980 records, as depicted in Table 7.

Table 7: Preprocess customer information

Attribute Name	Data Type
Customer Status	Character
Gender	Character
Nationality ID	Integer
ID Type	Character
Date of Birth	Date time
Client Type	Character
City	Character
MNO	Integer
Customer Profile	Character

- Transaction information consists of **564912** records, as shown in Table 8.

Table 8: Preprocess Transaction information

Attribute Name	Data Type
Transaction Date	Date time
Direction	Character
Status	Character
Total Amount	Integer
End Point Operation	Character
City	Character

- *Messages consist of **2632555** records, as shown in Table 9.*

Table 9: Preprocess message information

Attribute Name	Data Type
Message Type	Character
Operation	Character
Processing Status	Character
Processing Stamp	Date time

- Normalization: Transforming all string data into numerical values is crucial to enhance the efficiency and quality of subsequent processes.

Fuzzy Logic and association rules are applied to predict actions to improve company operations and provide exceptional customer services by utilizing connections between attributes and outcomes. We incorporate recovery rules in conjunction with the acquisition of segmented datasets. Relationships and regulations may involve one, two, or more attributes, whether from the same group (table) or distinct groups (tables), depending on the strength of associations between these attributes.

EXPERIMENTAL FINDINGS

In this Section, we present the outcomes of our experiments, offering insights into the data following comprehensive analysis and testing.

Our approach begins by assessing the strength of relationships between variables within each group, employing the Correlation Coefficient. Subsequently, we gauge the clustering rate through ANOVA statistical analysis, which evaluates whether the variability between groups exceeds the variability within the groups. This analysis relies on the F-value and P-value.

The F-value determines whether the variance between groups surpasses the variance within the groups. At the same time, the P-value indicates the presence of effects among groups, with the alpha (α) value set at 0.05. Distinct values are obtained for each group, and these results are presented individually.

Our data analysis and testing have uncovered numerous insights and expectations crucial for informed decision-making within the company. This encompasses identifying redundant data in the database, such as attributes related to customer activity, registration status, and customer status, which occupy storage space without significant benefits. Additionally, insights from customer behavior while using the application can be leveraged to enhance the application's settings and interface.

Figure 5 illustrates the distribution of customer data based on Gender. These findings underscore that many individuals have suspended their accounts or encountered rejections in their attempts to open accounts. This conclusion is drawn after conducting correlation analysis using R Studio. Understanding the underlying reasons is crucial to prevent a surge in such cases. Furthermore, it becomes evident that particular attention should be directed toward the female category, representing a significant portion of the surrounding community.

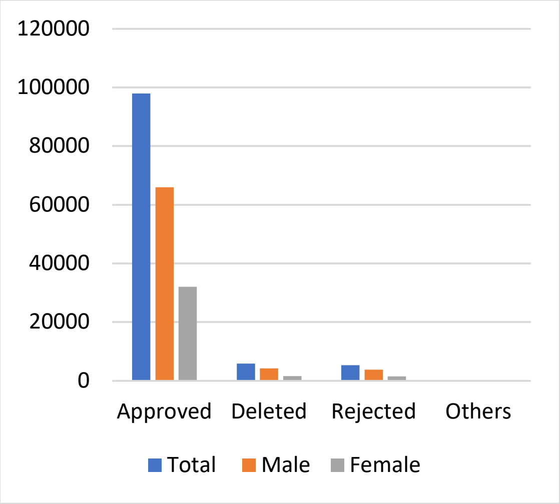

Figure 5: Correlation Based On Gender And Status.

Figure 6 illustrates data distribution across all cities based on Gender and customer status. The outcomes of the correlation analysis conducted with R Studio indicate that the male customer population surpasses the female population. This underscores the importance of directing efforts toward attracting more female users to broaden and diversify this user segment. Interestingly, the percentage of females who either close the application or face registration rejection surpasses that of males.

Considering the number of customers in each City and the transaction volume of each customer, we can make informed decisions regarding opening new branches. Specifically, we prioritize customers with higher transaction volumes to offer them agency stores in their respective cities. Additionally, we can introduce special offers in various stores, including supermarkets, pharmacies, cafés, malls, and more, based on transaction frequency.

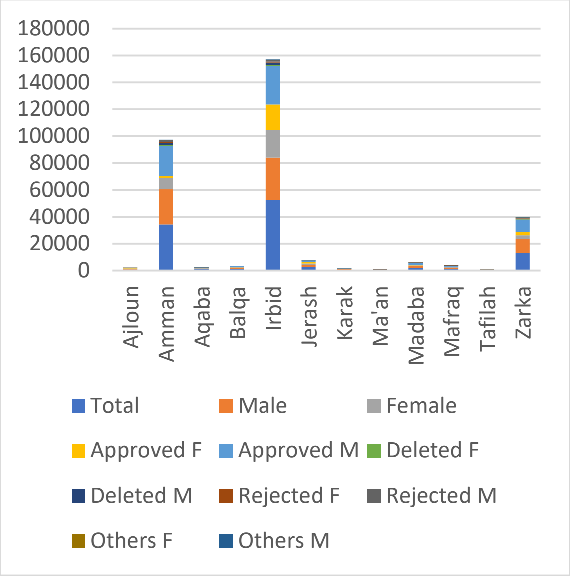

Figure 6: Correlation Based On Gender, Status, And City

Figure 7 displays the data distribution for all cities based on Gender. The results of the correlation analysis conducted using R Studio indicate that many application users are concentrated in three specific cities. Therefore, efforts should be directed towards expanding user presence in different cities.

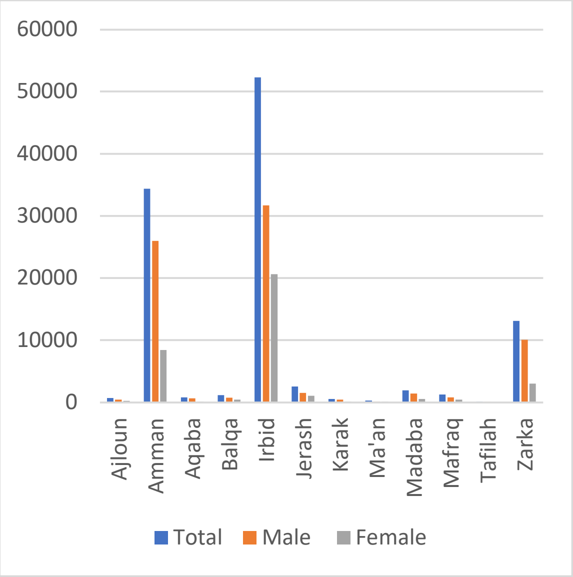

Figure 7: Correlation Based On Gender And City

Figure 8 illustrates the gender distribution within the telephone network. The results obtained through correlation analysis using R Studio reveal a notable concentration of customers affiliated with a specific telecommunications network. Notably, there might be a correlation between the mobile operator and the City, offering valuable insights for the service provider company to enhance services and introduce targeted offers in these areas.

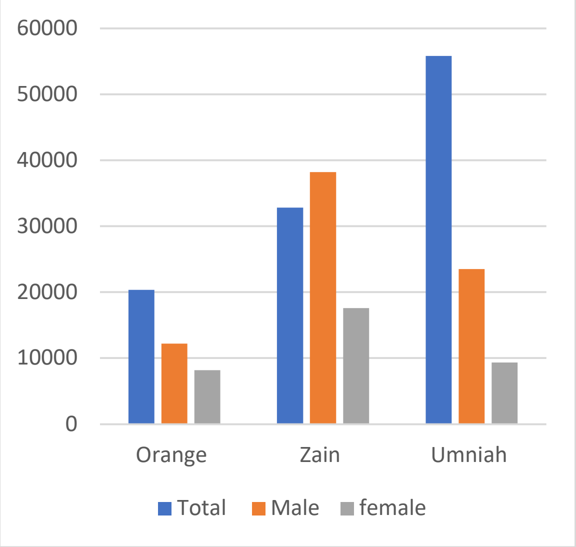

Figure 8: Correlation based on Gender and MNO

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 9 summarises the data by Gender, telephone network, and City. After applying correlation analysis using RStudio, the results emphasize a notable concentration of using a specific telecommunications network in a particular geographic area. This finding suggests an opportunity to provide enhanced services to this category of users and to target individuals from different communication networks in these areas.

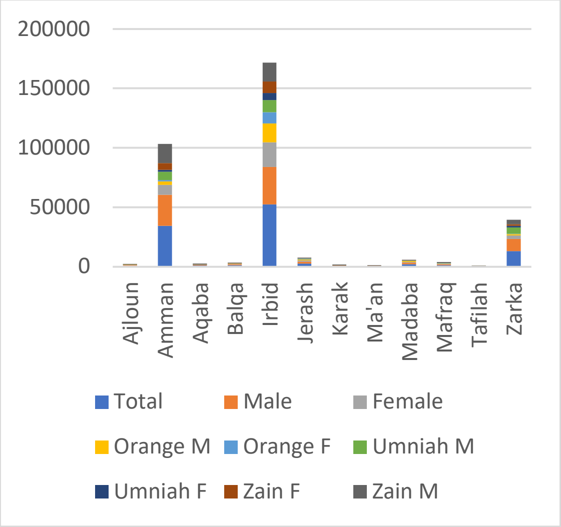

Figure 9: Correlation Based On Gender, MNO And City

Figure 10 depicts the distribution of data based on Gender and ID type. The results of correlation analysis using R Studio indicate a substantial percentage of application users with Jordanian identity. Hence, there is a need to focus on various social groups and segments that may contribute to increased transactions.

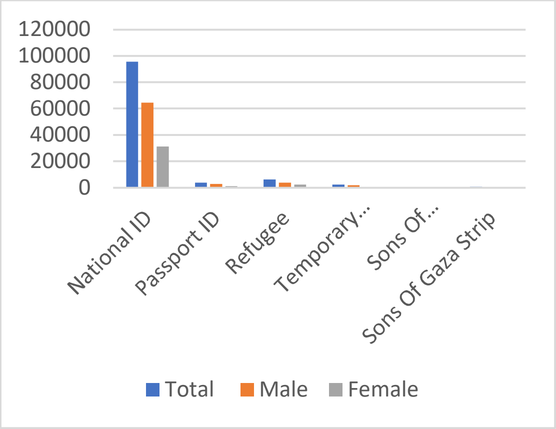

Figure 10: Correlation Based On Gender And ID Type

Figure 11 illustrates the data distribution based on the ID type and the City. The results and findings obtained from applying correlation analysis using R Studio underscore a distinction in the distribution of people based on population density and service focus. Thus, there should be an emphasis on the marketing process and service distribution in different cities.

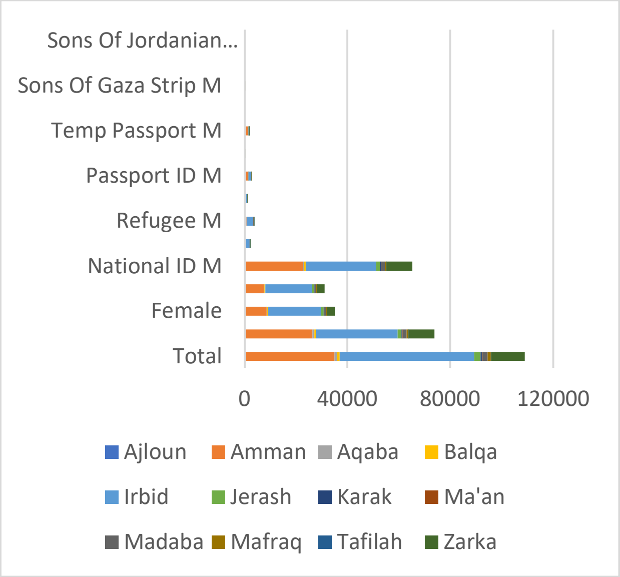

Figure 11: Correlation based on ID type and City.

Figure 12 presents the data distribution based on the transaction status and the City. The results and conclusions from applying correlation analysis using R Studio indicate that transactions exhibit a distribution and concentration in areas with higher density than in other regions. This sheds light on the transaction status within these areas.

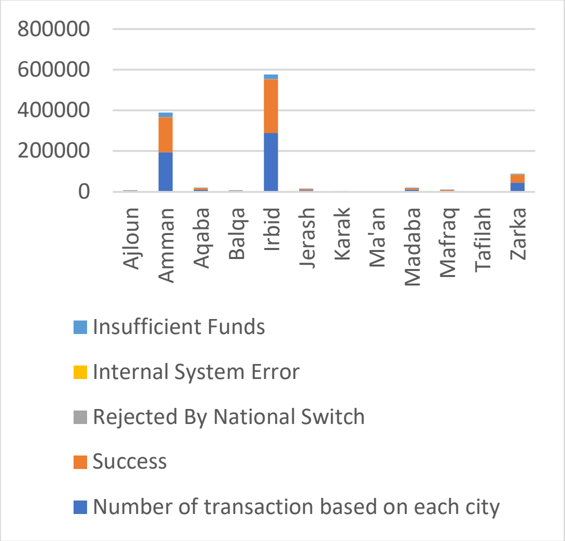

Figure 12: Correlation Based On City And Status

Figure 13 illustrates the data distribution based on the transaction direction and the City. The results and findings obtained through correlation analysis using R Studio emphasize a distinction in the direction of transactions, which is influenced by the services offered and the nature of the transaction used.

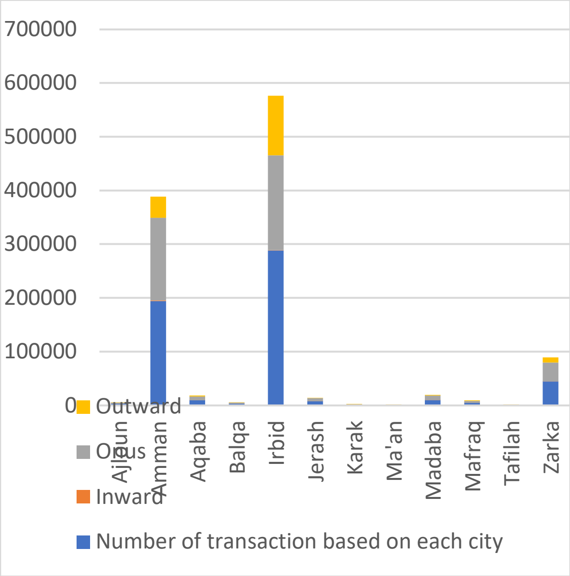

Figure 13: Correlation Based On City And Direction

Figure 14 displays the distribution of data based on the nature of the transaction and its state. The results and conclusions obtained from applying correlation analysis using R Studio suggest focusing on specific types of transactions over others, aligning with users' needs and interests. Therefore, emphasis should be placed on these types while efforts are made to increase the usage of other transaction types. Furthermore, there is a suspected transaction pattern in the application, such as a new account being opened followed by a customer depositing money—this behavior is deemed abnormal and warrants.

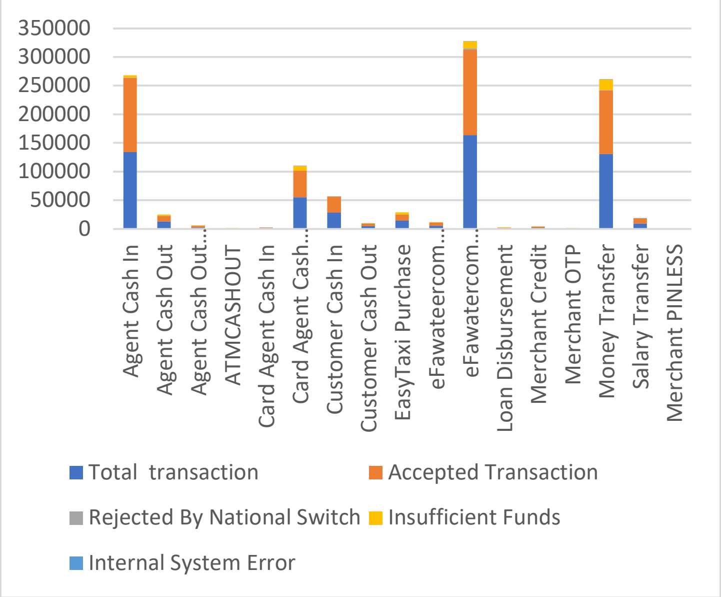

Figure 14: Correlation Based On Status Transaction.

5. CONCLUSIONS AND FUTURE WORK

Big Data has become ubiquitous in various industries and academic disciplines. Data originates from diverse sources across different domains, including industry, business, social networks, the Internet, health, finance, economics, and transportation.

This study introduced a novel approach for extracting meaningful insights from financial data—a strategy or framework designed to empower companies with high-quality information for precise decision-making. This framework comprises the following steps.

1. Data Collection Layer: The initial step of the process.

2. Data Preprocessing: This process converts semi-structured data into a structured format, preparing it for subsequent layers that involve

extracting rules and relationships between data (tables).

3. Clustering (Grouping) Data: The SVM clustering algorithm was utilized.

4. Analysis of data: Focus on acquiring meaningful insights.

Each of these layers encompasses components aimed at effectively addressing the challenges posed by big data.

Our approach leverages the capabilities of both the Spark and R frameworks and adequately addresses the three Vs of big data: Volume, Velocity, and Variety. Specifically, it utilizes SVM clustering and fuzzy logic algorithms, offering a unique approach for extracting meaningful and high-quality data, thus effectively representing semi-structured data.

In the experimental phase, we demonstrate the efficacy of our approach using a financial dataset in the mobile payment domain. We assessed the strength of the relationships between the variables within each group using the Correlation Coefficient. Additionally, we measured the clustering quality by employing ANOVA, which determined whether the variability between groups exceeded that within them. Using F-value and P-value allowed us to make informed decisions based on each group's "Rules."

Furthermore, our approach holds significant potential in the financial sector, where big data facilitates informed decision-making. It also encourages increased investment in the financial domain, particularly in mobile payment solutions. Several crucial aspects should be considered in future work. Our primary goal is to make this approach universally applicable to all financial companies operating in the mobile payment sector. It should be sufficiently versatile to manage various structured, unstructured, and semi-structured data types. In addition, it should be adaptable to dynamic and static data scenarios, enabling its application in a broad spectrum of use cases within the financial industry.

REFERENCES

[1] Analytics Vidhya https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ [on line 10/10/2023].

[2] Andrea, Marco, and Michele. (2015), What is big data? A consensual definition and a review of key research topics American Institute of Physics, doi:10.1063/1.4907823.

[3] Areen Al-Hgaish, Wael Alzyadat, Mohammad Al-Fayoumi, Aysh Alhroob, Ahmad Thunibat. (2019), Preserve Quality Medical Drug Data toward Meaningful Data Lake by Cluster International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878.

[4] Avita Katal, Mohammad Wazid and R. H. Goudar. (2013). Big Data: Issues, Challenges, Tools and Good Practices. 978-1-4799-0192-0/13/$31.00

[5] Khalid Adam Ismail Hammad, Mohammed Adam Ibrahim Fakharaldien, Jasni Mohamed Zain, Mazlina Abdul Majid. (2015), Big Data Analysis and Storage Proceedings of the 2015 International Conference on Operations Excellence and Service Engineering, Orlando, Florida, USA.

[6] Dhruba Borthakur, Joydeep Sen Sarma, and Jonathan Gray. (2011), Apache Hadoop Goes Real-time at Facebook.

[7] Ikhlas Almukahel, Wael Alzyadat, Mohamad Alfayomi. (2019), Hybrid Approach Using Fuzzy Logic and MapReduce to Achieve Meaningful Used Big Data International Journal of Engineering &Technology, 7 (4) 6997-7001.

[8] Indeed https://www.indeed.com/career-advice/career-development/correlation-definition-and-examples 2020.

[9] Jaseena K.U, Julie M. David. (2014), ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING Natarajan Meghanathan et al. (Eds): NeTCoM, CSIT, GRAPH-HOC, SPTM -pp. 131–140. © CS & IT-CSCP 2014, DOI: 10.5121/csit.2014.41311.

[10] Jasmine Zakir, Tom Seymour, and Kristi Berg. (2015), BIG DATA ANALYTICS, Issues in Information Systems, Volume 16, Issue II, pp. 81-90.

[11] Jin, S., J. Peng, and D. Xie. (2017), Towards MapReduce approach with dynamic fuzzy inference/interpolation for big data classification problems. IEEE 16th International Conference on Cognitive Informatics & Computing (ICCI* CC). IEEE.

[12] Mai Abdrabo, Mohammed Elmogy, Ghada Eltaweel, and Sherif Barakat. (2018), A Framework For Handling Big Data Dimensionality Based on Fuzzy-Rough Technique. Journal of Theoretical & Applied Information Technology. 96(4).

[13] NIST. (2018), Big Data Interoperability Framework: Volume 1, Definitions, NIST

Special Publication 1500-1r, version 2, National Institute of Standards and Technology, Gaithersburg, MD 20899 Information Technology Laboratory, Definitions and Taxonomies Subgroup, NIST Big Data Public Working Group (NBD-PWG) https://doi.org/10.6028/NIST.SP.1500-1r1

[14] Sam Madden. (2012), From Databases to Big Data, IEEE, Internet Computing.

[15] Sanjeev Dhawan and Sanjay Rathee. (2013), Big Data Analytics using Hadoop Components like Pig and Hive, American International Journal of Research in Science, Technology, Engineering & Mathematics, ISSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-3629, Available online at http://www.iasir.net.

[16] Sara del Río, Victoria López, José Manuel Benítez and Francisco Herrera. (2015), A MapReduce approach to address big data classification problems based on the fusion of linguistic fuzzy rules. International Journal of Computational Intelligence Systems, 8(3): p. 422-437.

[17] Parth Chandarana and M. Vijayalakshmi. (2014), Big Data Analytics Frameworks. International Conference on Circuits, Systems, Communication, and Information Technology Applications (CSCITA)

[18] Das, R. and I. Turkoglu. (2009), using the path analysis method, creating meaningful data from weblogs to improve a website's impressiveness. Expert Systems with Applications,36(3): p. 6635-6644.

[19] Educba https://www.educba.com/big-data-architecture/ [online 10/10/2023].

[20] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. (2014), Data Mining with Big Data, Transactions on Knowledge and Data Engineering, Vol. 26, No. 1. 1041-4347/, IEEE.

[21] Mohammed GH. AL Zamil, Samer Samarah. (2014), The Application of Semantic-based Classification on Big Data, International Conference on Information and Communication Systems (ICICS),978-1-4799-3023-4/14/$31.00.

[22] Potta Chakri, Saurabh Pratap, Lakshay, Sanjeeb Kumar Gouda. (2023), An exploratory data analysis approach for analyzing financial accounting data using machine learning,
Decision Analytics Journal, Volume 7.

[23] Leow K-R, Leow M-C, Ong L-Y. (2023), A New Big Data Processing Framework for the Online Roadshow. Big Data and Cognitive Computing. 7(3).

[24] Hasan, M.M., Popp, J. & Oláh, J. (2020), Current landscape and influence of big data on finance. J Big Data 7, 21.

[25] Imad Salah, Amani Alghareeb, Mohammad Aref Alshraideh, (2023), Navigating Ethics in Digital Humanities: A Deep Dive into Decision Distribution within Higher Education at the University of Jordan, International Journal on Recent and Innovation Trends in Computing and Communication, 11(9).

[26] Mohammad Alshraideh, Abeer Abdel-Jabbar Abu-Zayed, Martin Leiner, and Iyad Muhsen AlDajani, (2024), Beyond the Scoreboard: A Machine Learning Investigation of Online Games' Influence on Jordanian University Students' Grades, Applied Computational Intelligence and Soft Computing. https://doi.org/10.1155/2024/1337725.

[27] Nancy Shaar, Mohammad Alshraideh, Lara Shboul & Iyad AlDajani (2023) Decision support system (DSS) for traffic prediction and building a dynamic internet community using Netnography technology in the City of Amman, Journal of Experimental & Theoretical Artificial Intelligence, DOI: 10.1080/0952813X.2023 .2165716.

Table 4Describe the Data Type

Factors	Structured data	Semi-structured data	Unstructured data
Flexibility	It is dependent and less flexible.	It is more flexible than structured data but less flexible than unstructured data.	It is flexible, and there is an absence of a schema.
Transaction Management	Matured transactions and various concurrency techniques	The transaction is adapted from DBMS, not matured	There is no transaction management and no concurrency
Query performance	Structured queries allow complex joining	Queries over anonymous nodes are possible	An only textual query is possible
Technology	It is based on the relational database table	It is based on RDF and XML	This is based on character and library data

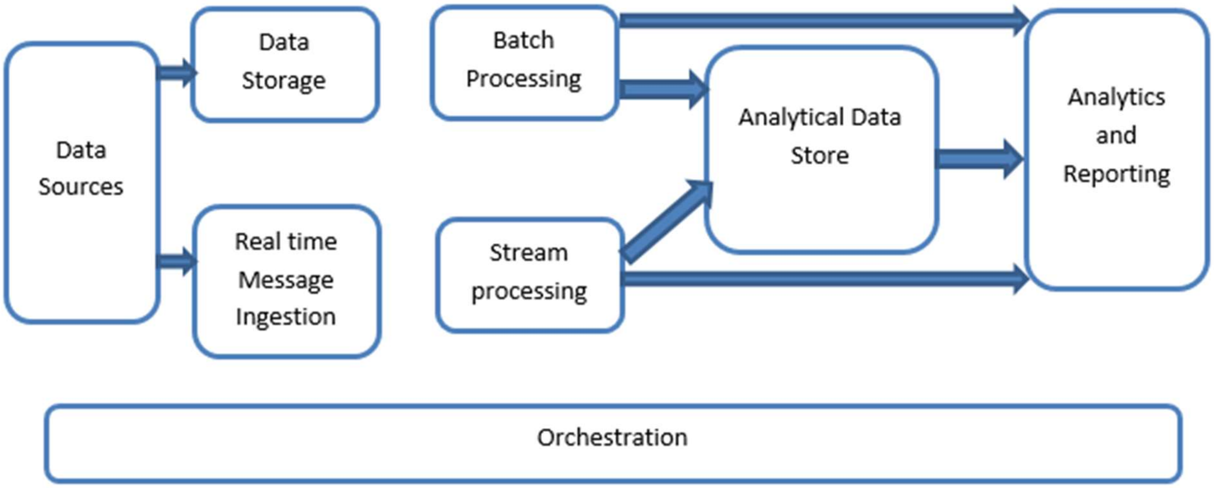

Figure 1: Explanation of Big Data Architecture

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

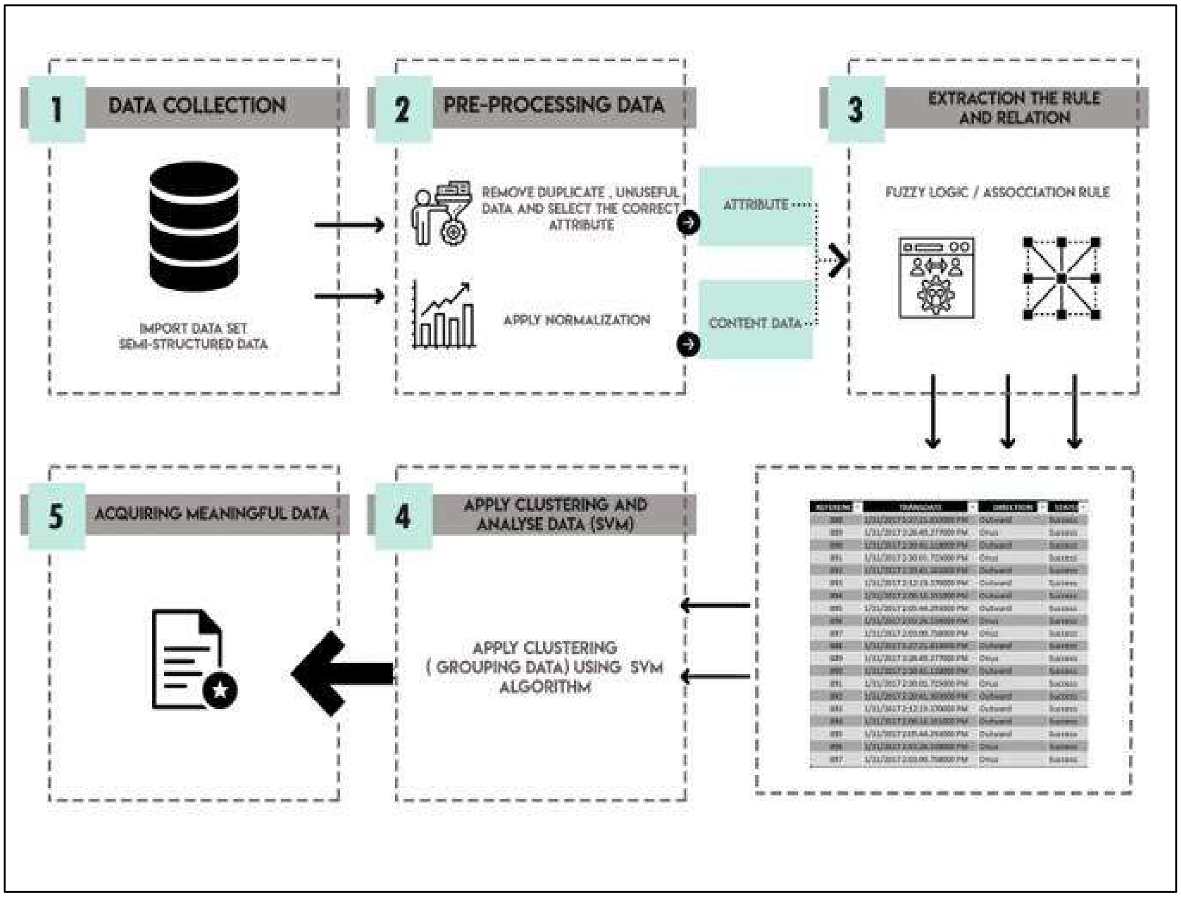
1
DATA COLLECTION
IMPORT DATA SET
SEMI-STRUCTURED DATA
2
PRE-PROCESSING DATA
REMOVE DUPLICATE , UNUSEFUL DATA AND SELECT THE CORRECT ATTRIBUTE
APPLY NORMALIZATION
ATTRIBUTE
CONTENT DATA
3
EXTRACTION THE RULE AND RELATION
FUZZY LOGIC / ASSOCCIATION RULE
4
APPLY CLUSTERING AND ANALYSE DATA (SVM)
APPLY CLUSTERING (GROUPING DATA) USING SVM ALGORITHM
5
ACQUIRING MEANINGFUL DATA

Figure 2: Proposed approach "Framework."

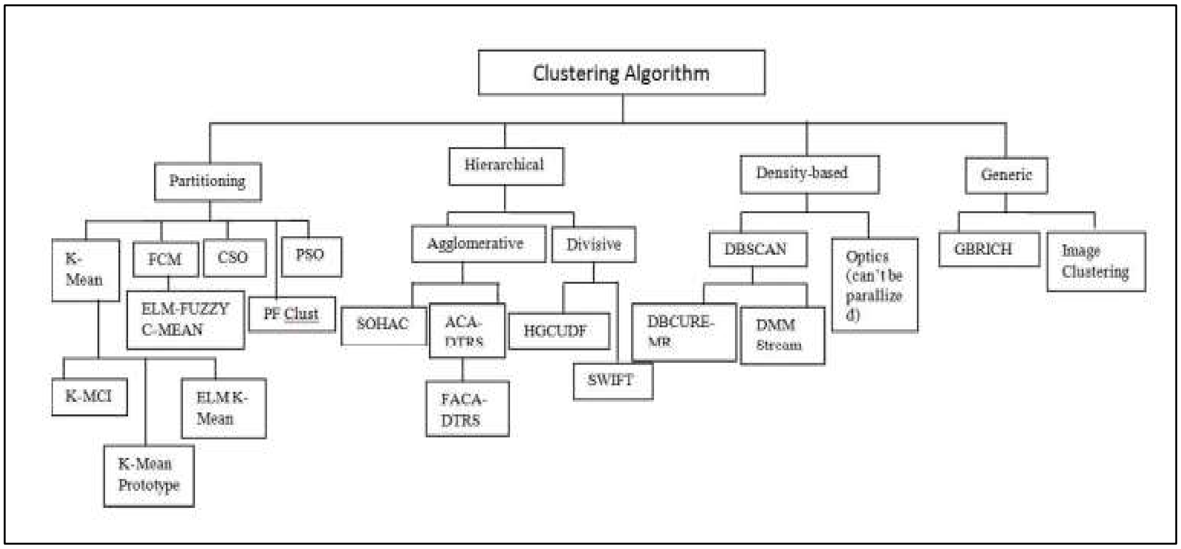

Figure 3: Type of Clustering Algorithm

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

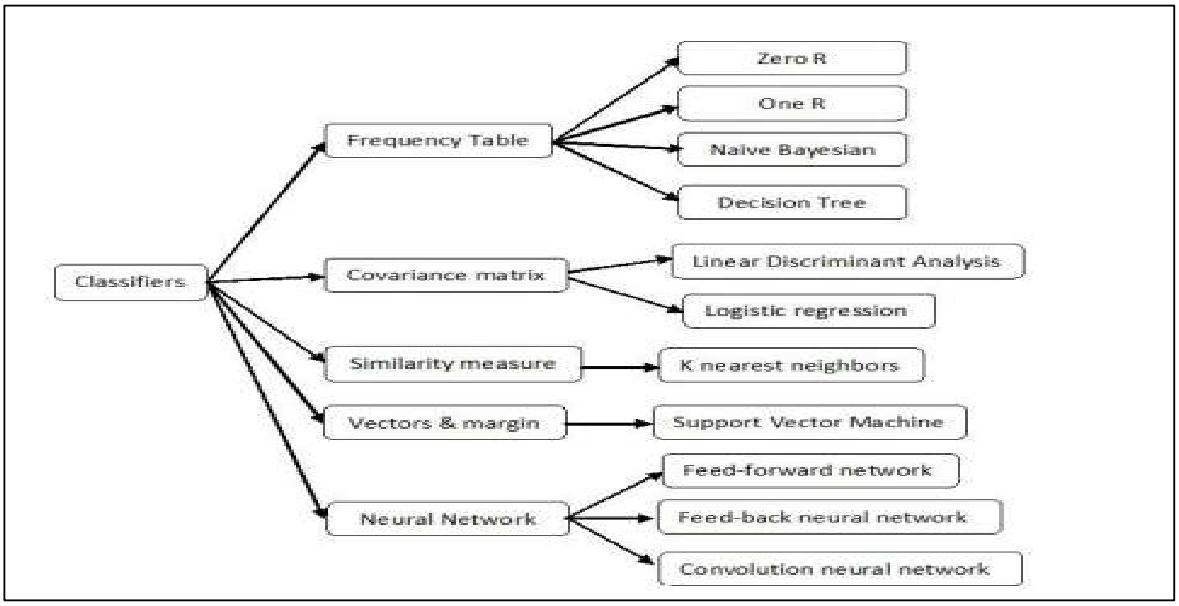

Figure 4: Type of Classification Algorithm