SALES DATA ANALYSIS USING MONGODB

A PROJECT REPORT

Submitted by

TANYA YADAV [RA2211031010112]

JIYA GAYAWER [RA2211031010129]

ANOUSHKA SHRIVASTAVA [RA2211031010135]

Under the Guidance of

Dr. ANGAYARKANNI S A

(Assistant Professor, NWC)

in partial fulfillment of the requirementsfor the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING

with specialization in INFORMATION TECHNOLOGY

DEPARTMENT OF NETWORKING AND COMMUNICATIONS

COLLEGE OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR- 603 203

NOVEMBER 2024

SRM

Department of Networking and Communications

SRM Institute of Science & Technology

Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/ Course	**: B.Tech / CSE-IT**
Student Names	**: Tanya Yadav, Jiya Gayawer, Anoushka Shrivastava**
Registration Numbers	**: RA2211031010112, RA2211031010129, RA2211031010135**
Title of Work	**: Sales Data Analysis Using Mongodb**

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly referenced / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

	Jiya Gayawer	Anoushka Shrivastava
TANYA YADAV	JIYA GAYAWER	ANOUSHKA SHRIVASTAVA

If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

SRM

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR – 603 203

BONAFIDE CERTIFICATE

Certified that 21CSC314P – Big Data Essentials mini-project report titled "**SALES DATA ANALYSIS USING MONGODB**" is the bonafide work of "**TANYA YADAV [RA2211031010112], JIYA GAYAWER [RA2211031010129], ANOUSHKA SHRIVASTAVA [RA2211031010135]**" who carried out the mini-project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Panel Reviewer I

12/11/24

SIGNATURE

Dr. Angayarkanni S A
Assistant Professor
Department of Networking
and Communications

Dept. of Networking and Communications
Kattankulathur
603 203
SRM IST

Panel Reviewer II

12/11/24

SIGNATURE

Dr. Sivamohan S
Assistant Professor
Department of Networking
and Communications

ABSTRACT

In today’s rapidly evolving retail landscape, businesses face mounting challenges in analyzing vast amounts of sales data to drive effective decision-making. Traditional data analysis methods often fall short, especially in capturing real-time customer buying patterns and preferences. This project addresses these challenges by developing a sophisticated system that leverages machine learning, big data technologies, and real-time customer interactions to comprehensively analyze sales data. The platform allows users to log in, write product reviews, and receive sentiment feedback—stored in SQL and processed through MongoDB for efficient analysis. The system delves into customer behavior and predicts future purchasing trends by analyzing product ratings and reviews. Advanced algorithms—such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and XGBoost—enhance predictive accuracy. Additionally, sentiment analysis on reviews provides an in-depth understanding of customer satisfaction. Visualization tools, including dashboards, pie charts, and graphs, offer clear insights into sales trends, highlighting top-performing products across various categories. By optimizing large-scale data processing and enabling personalized recommendations, this system supports businesses in improving inventory management and customer satisfaction, ultimately giving them a competitive edge in the retail market.

TABLE OF CONTENT

CHAPTER 1

INTRODUCTION

In today's highly competitive marketplace, firms need to undertake data analytics for the benefit of rich sales strategies and customer satisfaction. At the wake of e-commerce explosion, with real-time data accesses rapidly being made available, organizations are in an imperative position to harness advanced analytics techniques. The project is in developing a holistic [1] sales data analysis system that integrates machine learning algorithms with big data tools for actionable insights in business settings.

The major purpose of the system proposed is to be able to do real-time sales analysis to help organizations realize the behavior of clients, predict future trends, and thus optimize inventory management. The proposed [2] system would make use of the power of algorithms like Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and XGBoost to discover even those patterns in data that human analytical methods would not discern. The algorithms can handle large amounts of data and also provide reasonable predictive accuracy, making them suitable for sales forecasting.

Predictive analytics [3] combined with sentiment analysis of reviews related to products will be incorporated in the project, and fundamentally, this is important in assessing customer satisfaction. With sentiments derived from customers' perceptions about products from reviews, business operations and interest in better solutions are improved. Based on customers' sentiments, businesses can make decisions that are in line with consumers' taste, thus promoting loyalty towards products among the consumers.

This will further [4] assist the company in determining the highly selling products under the various categories through the sales volume analysis. This is well important to marketing, inventory control, and sales forecasting. Taking these points into consideration, businesses will know which to allocate their resources and capitalize on them, hence emerging trends.

To make the interpretation [5] of data easier, the project will incorporate any visualization capabilities, such as pie charts, graphs, and interactive dashboards, to make hard data digestible so that stakeholders can come up with quick data-driven decisions. To deal with enormous data, the system utilizes MongoDB, one of the most emerging big-data tools that is extensible and highly performance-driven, whereby large datasets can be processed efficiently so that the analysis remains timely and relevant.

This sales data analysis system will see fundamental transformation in the way business entities understand their customers, as it connects predictive analytics, sentiment analysis, and data visualization. Essentially, the business purpose of the project is to make shopping more satisfying and to develop more sales opportunities by giving businesses an advantage in competition.

This work is organized Section II as reviews related works. Section III outlines the proposed method, detailing its features and functionality. Results and discussion are found in Section IV, where the effectiveness of the system is analyzed. Finally, Section V concludes with key findings.

1.1 Overview

In today's fast-paced retail environment, businesses must adapt swiftly to shifting consumer behaviours and evolving market trends. The surge in available data has sparked a paradigm shift, transforming how retailers analyse and leverage information to make well-informed decisions. However, the core challenge lies in efficiently processing and interpreting massive volumes of sales data in real time. This project addresses this challenge by developing a comprehensive system that integrates machine learning and big data tools to analyse sales data with precision. By honing in on customer behaviour and predicting future purchasing trends based on product ratings and reviews, the system aims to optimize inventory management, enhance product recommendations, and boost overall customer satisfaction. Using advanced algorithms—Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and XGBoost—alongside sentiment analysis to assess customer satisfaction, this project introduces an innovative solution for retail analytics.

1.2 Background

The retail sector is experiencing a profound transformation, fuelled by rapid technological advancements and evolving consumer expectations. The rise of e-commerce and mobile shopping has generated an immense volume of sales data, offering retailers invaluable insights into customer preferences and opportunities to optimize their offerings. However, traditional data analysis methods often lack the flexibility and depth needed to respond effectively to these shifts. Machine learning and big data analytics have become essential tools in overcoming these challenges. By deploying algorithms that identify and learn from data patterns, retailers can predict customer behaviour, personalize marketing efforts, and manage inventory more effectively. Adding sentiment analysis to this approach enhances its impact by allowing businesses to measure customer satisfaction and refine their strategies accordingly.

1.3 Problem Statement

Despite significant advancements in data analytics, many retailers continue to struggle with effectively analysing and utilizing sales data. Traditional data analysis methods are often inefficient and unable to capture the real-time nuances of customer behaviour, leading to missed insights into buying patterns. This results in suboptimal inventory management, decreased customer satisfaction, and a lack of personalized product recommendations. There is a growing need for a more robust system that integrates machine learning and big data technologies to facilitate comprehensive sales data analysis. Such a system would enable retailers to make timely, data-driven decisions that align with customer preferences and market demands, driving business growth.

1.4 Objectives

The primary objectives of this project are to:

1. **Develop a comprehensive system** that integrates machine learning and big data technologies for effective analysis of retail sales data.

2. **Implement advanced machine learning algorithms**, including Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and XGBoost, to predict customer behaviour and future purchasing trends based on historical sales data.

3. **Integrate sentiment analysis** to assess customer sentiment through product ratings and reviews, providing deeper insights into customer satisfaction.

4. **Visualize sales trends** using interactive dashboards, pie charts, and graphs, enabling intuitive understanding and detailed analysis of sales data.

5. **Enhance decision-making processes** for inventory management and personalized product recommendations, ultimately boosting customer satisfaction and driving sales growth.

CHAPTER 2

LITERATURE SURVEY

2.1 General

The rapid growth of Internet of Things (IoT) technologies has profoundly impacted data management and analysis across various industries, bringing both opportunities and challenges. This literature survey addresses key issues such as data pricing, sales prediction, visual analytics, and time series classification, which remain pivotal in both research and practical applications. Recent studies propose innovative solutions to these challenges, including simulation-based pricing models designed to navigate the complexities of IoT data markets and the application of graph neural networks to optimize constraints in sales prediction.

In addition, model-driven engineering approaches are explored to enhance visual analytics systems, ensuring they meet real-time decision-making needs. The automation of insights extraction from e-commerce reviews, utilizing advanced clustering techniques, highlights a growing trend to improve consumer sentiment analysis. Furthermore, the classification of multivariate time series data and the development of robust neural networks are emphasized as critical for achieving accurate data analysis. Collectively, these contributions demonstrate the evolving landscape of data-driven methodologies, underscoring their role in facilitating quality decision-making across diverse applications.

2.2 Literature Survey

This has greatly enhanced data management and analysis across Internet of Things devices and various domains. However, areas such as data pricing, sales prediction, visual analytics, and time series classification remain critical issues in the research and practice arena. It discusses recent novel approaches to overcome such challenges, including simulation-based pricing models, graph neural networks for the optimization of constraints, model-driven engineering for visual analytics, and advanced classification techniques for variable-length time series. This survey will synthesize such contributions to outline and highlight the changing landscape of data-driven methodologies and how these affect quality decisions in various applications.

Sensor clouds rely on data [6] from many devices with IoT connection as a means of linking sensors and users. There are numerous proposals for market frameworks which have been focused on the different stakeholders and their interests. However, the duplicability of the IoT data increases the challenges in the design of a natural pricing scheme. This work finally presents a seller-consumer competition-based pricing scheme that reflects the interplay between seller competition and consumer demand. These findings are obtained through comprehensive simulations of the market, which illustrate the desirable properties of such pricing in the IoT data market.

The business intelligence of a firm primarily depends on the correct prediction of sales. It further dictates the decisions regarding the production [7] levels and supply chain planning. Many researchers have this problem of obtaining optimum forecasts of commodity sales in the presence of explicit constraints. This work introduces a model for predicting based on a combination of constraint graphs and store graphs. A graph convolutional neural network is exploited to capture temporal features, which are then optimized for the predictive process. The developed model demonstrates phenomenal accuracy for comparative performances against classical approaches.

Data consolidation from [8] heterogeneous sources and the generation of user-specific visualizations are required for visual analytics systems that may facilitate data-driven decisions. However, the support provided by existing solutions may not satisfactorily meet these requirements. This work outlines model-driven engineering as a direction for the design of visual analytics systems. A Domain-Specific Modeling Language (DSML) named ModelViz is developed for consumer goods supply chain applications. Quantitative evaluations demonstrate that this approach properly meets the users' needs and permits promising directions for future design.

E-commerce reviews provide insight into the opinion of consumers. However [9], it is practically impossible to read such huge quantities of reviews. The saving of time and cost can be achieved through the automation of extracting useful insights from reviews. In this research, clustering algorithms have been used to identify related products, pros and cons, and trends. A dataset has been constructed on multiple products and relevant online platforms to support the research. It is along with these perceptions that the use of sophisticated clustering techniques shall increase further to comprehend consumer mood as an effective marketing strategy.

Multivariate time series classification automatically takes real-world data analysis tasks [10], with ROCKET having recently been established as an effective algorithm for accurate classification. However, the common assumption of equal-length time series does not hold in practice. This work explores preprocessing pipelines to variable-length time series that need classification. Three methods are analysed-truncation, padding, and forecasting-and padding is recommended as most effective. Incident detection in cash transactions also serves to attract attention to additional challenges presented by imbalanced data and variable length.

For the last decade, deep neural networks have outperformed traditional models for [11] machine learning on most supervised tasks. Most of the models are optimized with Cross-Entropy function, which has poor margins and is not stable. This work proposes a supervised contrastive learning framework that boosts the inter-class separability and the intra-class compactness of textual representations. It uses a novel contrastive loss and also develops a method for selecting hard negatives during training. Extensive experiments demonstrate the method outperforms several competing approaches on large-scale text classification benchmarks.

This work follows up on recent works on social robots in retail application [12], especially on using a single robot as a recommender for the products. Within this work, we hypothesize and explore a wingman-leader recommendation strategy where the wingman robot recommends its leader to increase sales. The outcome shows that it attracts many customers to the leader robot, which, therefore improves on the sales performance. Using two robots with such a strategy makes it more effective than no setups at all. In this regard, this work contributes to understanding cooperative strategies in robotic product recommendations.

Distribution of VAT value-added tax in China follows the origin principle. This tendency [13] exacerbates the difference in regional tax revenue due to the retail boom in online sales. This work examines big e-commerce transaction data in the light of its ability to pinpoint regional imbalances in retailing and consumption. A game theory model of the origin principle explained how it creates tax competition between regions. The work suggests some policies to solve these inequalities, and it argues that the imposition of a destination principle can minimize tax inequalities. Further relief in revenue inequality can be seen in changes of the portion of distribution between local and central governments.

The rise in smart environments has increased the amount of generated data [14], which calls for effective management solutions. Blockchain technologies are a safe and transparent alternative for processing data but come with more critical security problems. Motivation: The motivation to start this work is to explore how artificial techniques may be combined for anomaly detection in blockchain networks. This work proposes a framework that explores combining blockchain with AI on security matters. Major challenges and trends in improving the security of blockchain technologies in smart environments were obtained from research findings.

This work [15] summarizes the main sales models of the automobile sales industry and analyzes the internal and external environment of the automobile sales industry using the five forces model. Furthermore, the work collects data through questionnaires and conducts research on automobile marketing by using principal component analysis and SICAS model, which provides a theoretical basis for the automobile sales model under the new retail.

This research [16] develops auxiliary evaluation criteria that could be helpful in getting a more accurate data review based on contemporary categorization methods. Using the after-sales energy supplement context in the renewable energy sector as an example, the use of different criteria is established through the expert rating method. Thereafter, successive adjustments are made by consistency and concentration checks using the Kendall coefficient, thus leading to determining the weight of each criterion. This process provides a reference framework for data classification and grading concerning the after-sales energy supplement context and the broader automotive industry. This article [17] explains the information systems of car OEMs regarding after-sales and conducts a comprehensive analysis of the primary data applications used in spare parts planning.

The study [18] aimed to design a model on salesperson performance by means of the clustering of sales data. The study applied the framework of CRISP-DM. As such, the proposed model is connected with the existing sales order database. The model fetches multidimensional characteristics and classifies data labels from the database. The multidimensional features were obtained by using the Kohonen SOM clustering evaluation, which resulted in a quantization error of 0.95 and a topographic error of 0.13.

This research [19] uses the algorithms of KNN, RF, and SVM as core classifiers, which are then combined using a soft voting mechanism to create the KNN-RF-SVM ensemble model. The experiments led to results signifying that the integrated KNN-RF-SVM model outperforms the standalone method models in terms of accuracy, precision, recall rate, and F1 score.

Consequently, this study aims [20] to enhance the management service level of electricity sales channels by focusing on recent developments in electric power marketing within a specific region. Employing extensive datasets to assess the operational management of electricity sales channels at both national and global levels, a model for evaluating system dynamics is developed. This approach facilitates a detailed study of the business environment associated with electricity sales modes, and it also provides an integrated analysis of the assessment results, which leads to the formulation of strategic measures.

The work [21] propounds the concept of data mining, further develops the random forest algorithm in this regard and proposes an avenue that incorporates Bayesian optimization as well as time series segmentation to improve the random forest model, thereby increasing predictive accuracy. This research established a sales forecasting model tailored for mobile e-commerce, utilizing an improved random forest algorithm. The analysis of the results from the prediction model reveals that the forecasting error is more pronounced for categories such as food and apparel, which are notably affected by various activities, while it is diminished for tools, whose sales experience less influence from such activities.

This study [22] first preprocesses the original data and then predicts car spare parts sales by carrying out time series analysis coupled with regression decision tree analysis. The findings from the experiment demonstrate that the time series analysis technique shows a significant level of correlation in line fitting between the forecasted and actual sales figures for auto spare parts, achieving an average prediction accuracy of about 90.21%. This indicates its appropriateness for estimating the yearly sales volume of auto spare parts.

The software developed for this objective is currently undergoing continuous improvements. SPOT represents an application specifically designed for the production of sales forecasts [23], based on unique input variables, which include a retail location, its conditions, unique products to be merchandised, and their corresponding product classes or subclasses that support decision-making within various sectors. The output is in the form of graphical displays of the chosen data set along with predictions based on time-series analysis. Enhanced sales forecasts are exported in the form of a tabular numeric file. Using data from Walmart, the whole process will be illustrated, but this can be applied to other relevant application domains as well.

This study [24] uses PCA to reduce the dimension of the input features for the LSTM network to preserve their essential information while simultaneously exploiting the strengths of both the XGBoost model and the LSTM neural network. After the removal of noisy data, using historical entity data, a sales forecasting model that incorporates PCA, LSTM, and XGBoost is built. Comparative experiments are designed and run with one chosen model compared against two established time series forecasting models. Experiments showed that under the same scenario, the PCA_LSTM-XGBoost combined predictive model could decrease the MAPE by at least 8% in commodity sales volume forecasting compared to alternative models, showing superior accuracy and enhanced applicability.

2.3 Summary

The literature survey highlights a diverse range of studies addressing critical challenges in data management within IoT and retail environments. A prominent issue discussed is the need for effective pricing mechanisms in IoT data markets, where the duplicability of data complicates the design of natural pricing schemes. Innovations such as seller-consumer competition-based pricing models offer insights into market dynamics through simulations, helping balance seller competition with consumer demand.

In the field of sales prediction, researchers are integrating constraint graphs with graph convolutional neural networks to enhance forecasting accuracy. These models tackle explicit constraints impacting sales predictions and significantly outperform classical approaches. Furthermore, model-driven engineering approaches in visual analytics are advancing, facilitating user-specific data visualizations that enable business-driven, data-informed decisions.

The automation of extracting insights from e-commerce reviews through clustering algorithms reflects a growing focus on efficiently analysing consumer sentiment and trends, which can inform marketing strategies. Additionally, advancements in multivariate time series classification, particularly using the ROCKET algorithm, underscore the importance of addressing variable-length time series data, with preprocessing techniques such as padding shown to improve classification accuracy.

Recent developments in deep learning, particularly supervised contrastive learning, are providing new methods to enhance text classification accuracy, outperforming traditional models. Research into social robots for retail applications demonstrates the potential of cooperative recommendation strategies, where robots work together to enhance sales performance.

The literature also addresses regional imbalances in e-commerce tax revenue through game theory models, highlighting the need for equitable policies in response to the growing e-commerce market. In addition, the integration of blockchain with AI for enhanced security in smart environments offers a promising direction for managing the increasing complexity of data.

Overall, this literature survey synthesizes recent advancements in data-driven methodologies, underscoring their impact on improving decision-making in IoT and retail sectors. It emphasizes the importance of these innovations in shaping future research and offering practical solutions to contemporary challenges in these domains.

CHAPTER 3

PROPOSED METHODOLOGY

The methodology for this project focuses on real-time sales data analysis, incorporating multiple stages, from data collection to visualization. The process begins with gathering data from various sources, such as e-commerce platforms and customer feedback systems. This is followed by data preprocessing, model development, sentiment analysis, and predictive analytics. Finally, the results are visualized to facilitate decision-making. This comprehensive approach ensures businesses gain a holistic understanding of customer behavior, enabling them to make data-driven decisions that improve customer satisfaction and overall performance.

3.1 Module Description

3.1.1 Data Collection:

This module is responsible for gathering sales data from multiple sources. These include e-commerce platforms, where sales transactions are recorded, and customer feedback systems, where product reviews are captured. By combining these datasets, the system ensures an effective analysis of both sales performance and customer sentiment. This integrated dataset serves as the foundation for further analysis. Additionally, to test and evaluate the performance of our models, we used the **Amazon Dataset Review** dataset (https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/) , which provides extensive product reviews and ratings, allowing for robust model testing and evaluation.

3.1.2 Data Preprocessing:

Data preprocessing ensures the quality and usability of the collected data. The preprocessing steps include data cleaning to remove duplicates and handle missing values. It also standardizes the data format for consistency. For text data, preprocessing involves tokenization and removing stop words before running sentiment analysis on customer reviews. This ensures that the data is clean, reliable, and ready for deeper analysis.

3.1.3 Model Development with Machine Learning:

The core element of the methodology is developing models in machine learning based on an analysis of sales data. Algorithms applied for forecasting behavior by purchasing a customer and the best-selling products are Support Vector Machines, Decision Trees, Random Forest, and XGBoost. Models are trained on historical data, including the Amazon dataset, and their accuracy and performance are evaluated to determine the best-performing model for predictions.

3.1.4 Sentiment Analysis:

There is also the analysis of customer review sentiment for determining the overall level of satisfaction. Reviews can be classified as being either positive, negative, or neutral through techniques used in NLP. Such an analysis often brings to the surface customers' opinions that allow business entities to customize their products as well as services.

3.1.5 Predictive Analysis:

This project uses predictive analytics to predict subsequent purchases by customers based on ratings and reviews. A system, using the trained machine learning model, could tell patterns of customer behavior and predict their next probable purchase, making it invaluable knowledge that can be used for inventory management and marketing strategy.

3.1.6 Data Visualization:

Data visualization techniques are used to make analytical results easier to understand. It produces pie charts, graphs, and interactive dashboards for key insights in an accessible format. A unified set of information from real-time sales data and customer sentiment provides all stakeholders with the best possible information so that they can take appropriate decisions.

3.1.7 Big Data Tools Implementation:

The methodology integrates MongoDB, which is a powerful NoSQL database useful for efficiently processing huge datasets. This option allows storing and retrieving data on an enormous scale; this will ensure that the system does not compromise performance while handling massive sales transactions and customer reviews. The integration of big data tools with machine learning has helped boost the overall effectiveness of the sales analysis system.

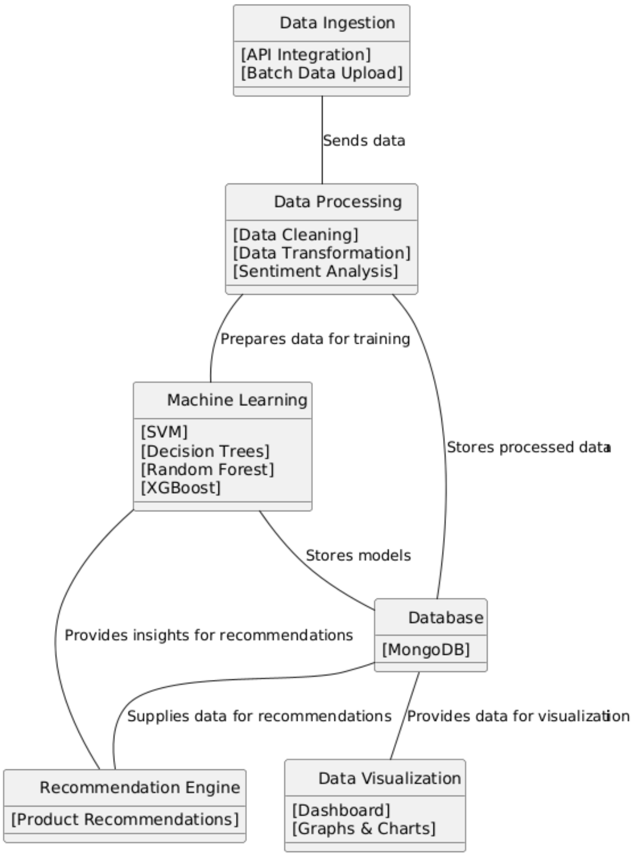


Fig. 3.1: Architecture Diagram

In fig. 3.1, the Architecture Diagram illustrates flow of Data in Real Time Sales Analysis Architecture In the data ingestion, it comes via APIs/Batch Upload fed to the processing layer where data cleaning, transformation, and sentiment analysis are performed. After processing, it sends data to some machine learning models like SVM, Decision Trees, Random Forest, and XGBoost to make predictions. Both processed data and machine learning models are stored in MongoDB. The data fetched from the database are used both to visualize on dashboards and for the purpose of generating recommendations about products. Other insights gained from the various machine learning models are applied to the recommendation engine to enhance customer experience.

3.2 System Requirements

3.2.1 HARDWARE REQUIREMENTS

- HDD: >90GB
- Processor: >Pentium IV 2.4GHz

- System Type: 32bit / 64 bit
- RAM: >2GB
- OS: Windows 7/8/8.1/10

3.2.2 SOFTWARE REQUIREMENTS

- Tool: MATLAB
- Toolbox: Image Processing Toolbox

3.3 Software Description

MATLAB is a flexible tool for data mining, though it is less commonly used standalone compared to tools like Clementine or Weka. It is often combined with other software, but its versatility allows for powerful data analysis and model building. Despite being ranked 10th in a 2016 data mining tools poll (**Fig 3.2**), MATLAB's potential remains underutilized due to its proprietary nature. However, open-source toolboxes and scripts can enhance its capabilities.

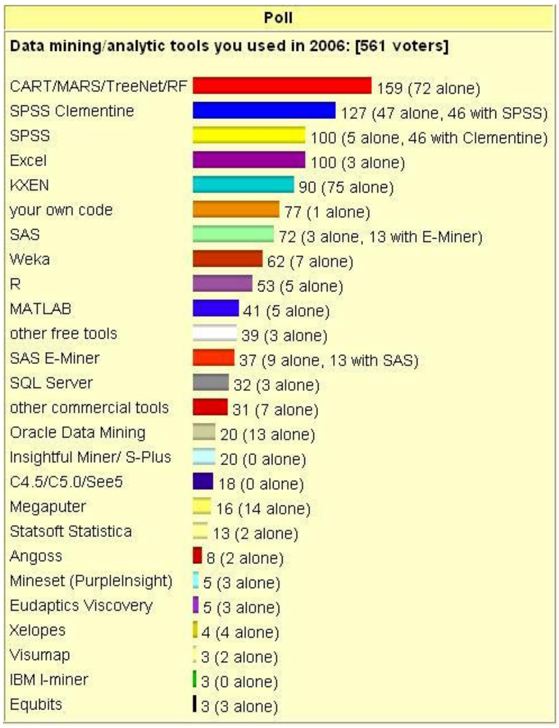


Figure 3.2: 2016 Data Mining Tools Poll 1138 Votes MATLAB Ranks 10th with 5% of the votes

Table 3.1 shows that decision trees, clustering, and neural networks were the most popular methods from 2013 to 2016, aligning with the techniques explored in this project. **Table 3.2** indicates

MATLAB's consistent presence in data mining, ranking between 7th and 15th from 2010 to 2016. This highlights its potential, even though it hasn't become the top choice for data mining.

Method	**2013**	**2014**	**2015**	**2016**
Decision tree	Rank:1 (15%)	Rank:1 (15%)	Rank:1 (16%)	Rank:1 (13%)
Clustering	Rank:2 (11%)	Rank:2 (11%)	Rank:3 (10%)	Rank:2 (12%)
Neural nets	Rank:5 (8%)	Rank:4 (8%)	Rank:5 (8%)	Rank:6 (7%)
Association rules	Rank:6 (7%)	Rank:7 (4%)	Rank:4 (8%)	Rank:7 (6%)

Table 3.1: Polls of trendy Data Mining Methods 2013-2016

MATLAB	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Rank	∞	7.0	7.0	14.0	9.0	15.0	10.0
Percentage	N/a	5%	5%	3%	2%	2%	5%

Table 3.2: celebrity of MATLAB in Data Mining 2010-2016

Integrating various data mining tools in MATLAB enables a more comprehensive approach, allowing it to serve as a stand-alone platform for sophisticated analysis. This case study illustrates how MATLAB can be effectively used for data mining tasks and suggests that a dedicated data mining toolbox could enhance its functionality.

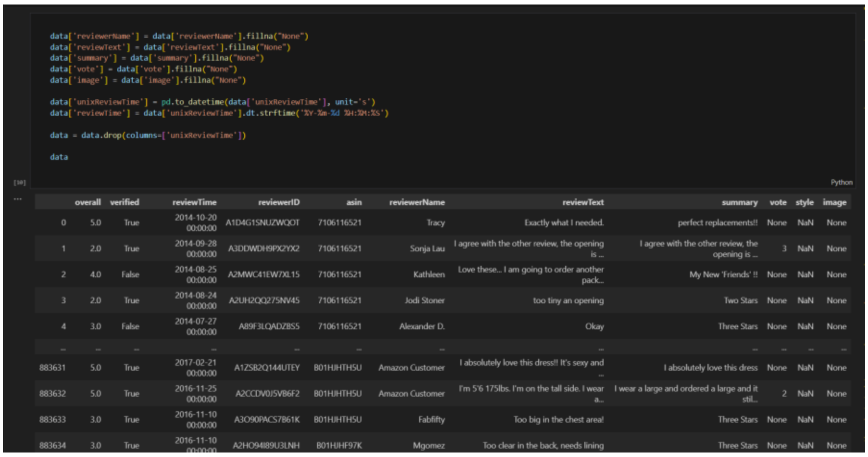
CHAPTER 4

IMPLEMENTATION

The implementation of the proposed technique involves a collaborative approach, with each team member assigned specific roles to ensure the effective handling of sales data and product reviews for predictive analysis.

4.1 Data Preprocessing

Data preprocessing included several key steps to prepare the dataset for analysis. First, data cleaning was conducted to remove duplicates, address missing values, and standardize data formats for consistency. Outliers were also identified and handled to prevent skewed results in the model.



```python
data['reviewerName'] = data['reviewerName'].fillna("None")
data['reviewText'] = data['reviewText'].fillna("None")
data['summary'] = data['summary'].fillna("None")
data['vote'] = data['vote'].fillna("None")
data['image'] = data['image'].fillna("None")

data['unixReviewTime'] = pd.to_datetime(data['unixReviewTime'], unit='s')
data['reviewTime'] = data['unixReviewTime'].dt.strftime('%Y-%m-%d %H:%M:%S')

data = data.drop(columns=['unixReviewTime'])

data
```

	overall	verified	reviewTime	reviewerID	asin	reviewerName	reviewText	summary	vote	style	image
0	5.0	True	2014-10-20 00:00:00	A1D4G1SNUZWQOT	7106116521	Tracy	Exactly what I needed.	perfect replacements!!	None	NaN	None
1	2.0	True	2014-09-28 00:00:00	A3DDWDH9PX2YX2	7106116521	Sonja Lau	I agree with the other review, the opening is ...	I agree with the other review, the opening is ...	3	NaN	None
2	4.0	False	2014-08-25 00:00:00	A2MWC41EW7XL15	7106116521	Kathleen	Love these... I am going to order another pack...	My New 'Friends' !!	None	NaN	None
3	2.0	True	2014-08-24 00:00:00	A2UH2QQ275NV45	7106116521	Jodi Stoner	too tiny an opening	Two Stars	None	NaN	None
4	3.0	False	2014-07-27 00:00:00	A89F3LQADZBS5	7106116521	Alexander D.	Okay	Three Stars	None	NaN	None
...	...	...	...	...	...	...	...	...	...	...	...
883631	5.0	True	2017-02-21 00:00:00	A1ZSB2Q144UTEY	B01HJHTH5U	Amazon Customer	I absolutely love this dress!! It's sexy and ...	I absolutely love this dress	None	NaN	None
883632	5.0	True	2016-11-25 00:00:00	A2CCDV0J5VB6F2	B01HJHTH5U	Amazon Customer	I'm 5'6 175lbs. I'm on the tall side. I wear a...	I wear a large and ordered a large and it stil...	2	NaN	None
883633	3.0	True	2016-11-10 00:00:00	A3O90PACS7B61K	B01HJHTH5U	Fabfifty	Too big in the chest area!	Three Stars	None	NaN	None
883634	3.0	True	2016-11-10 00:00:00	A2HO94I89U3LNH	B01HJHF97K	Mgomez	Too clear in the back, needs lining	Three Stars	None	NaN	None

Fig. 4.1: Missing values are nulled

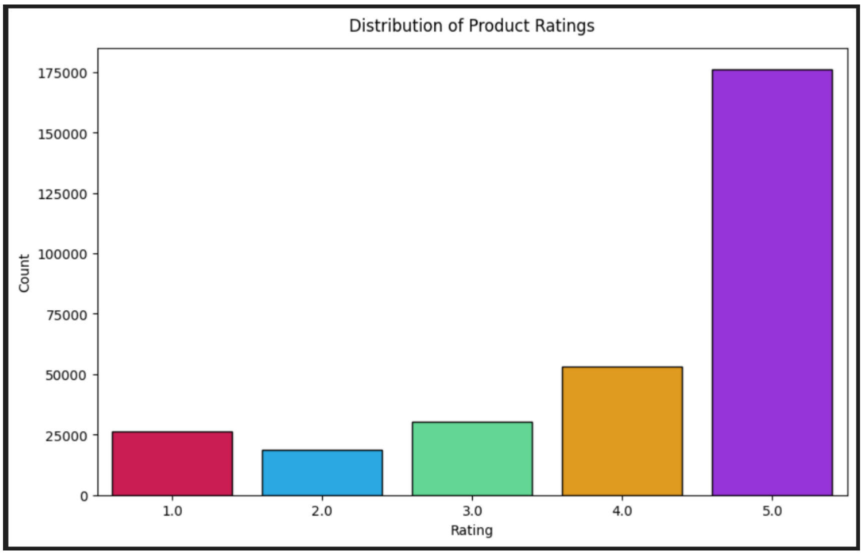
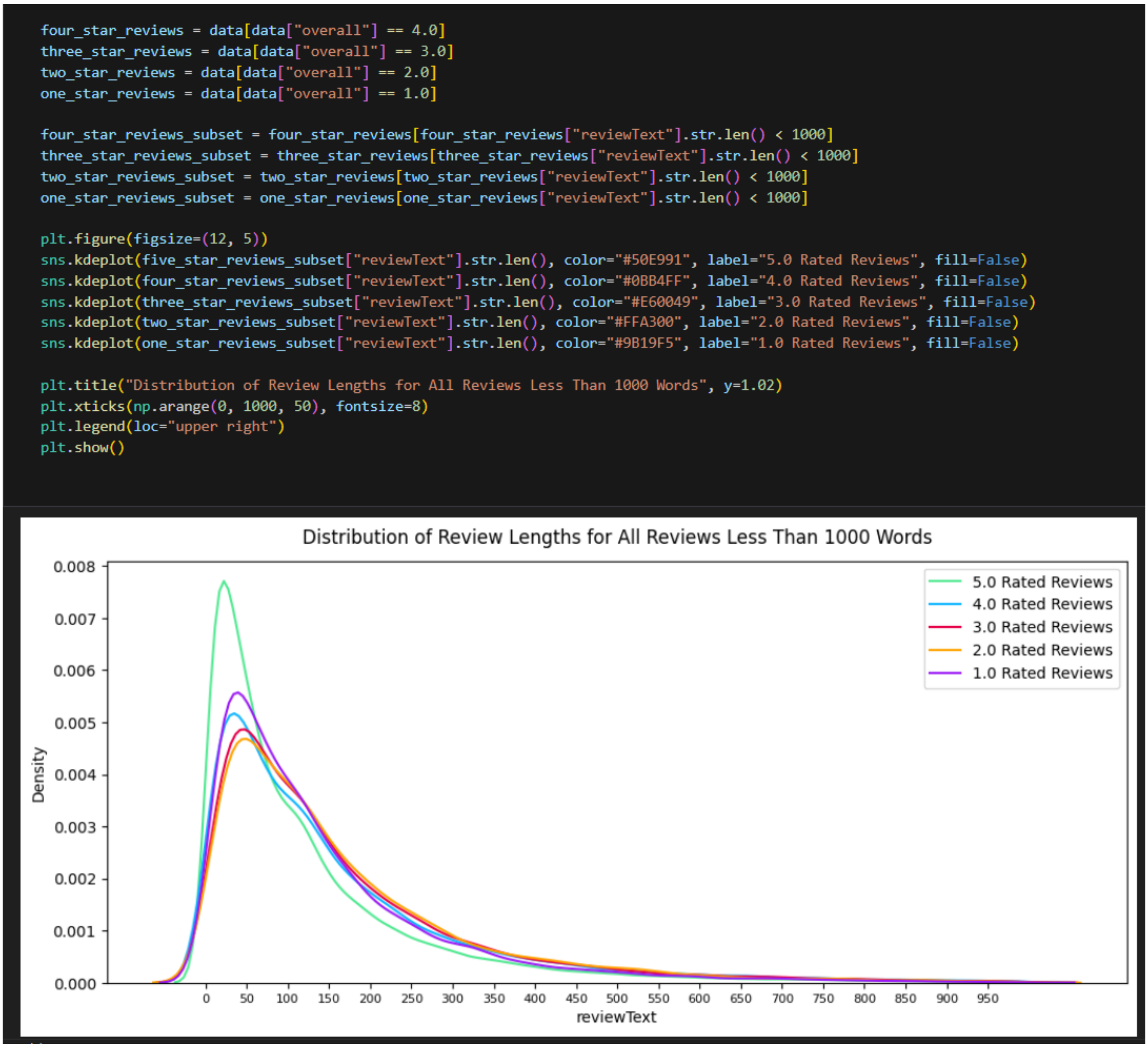


Fig. 4.2: Visualizes the distribution of product ratings

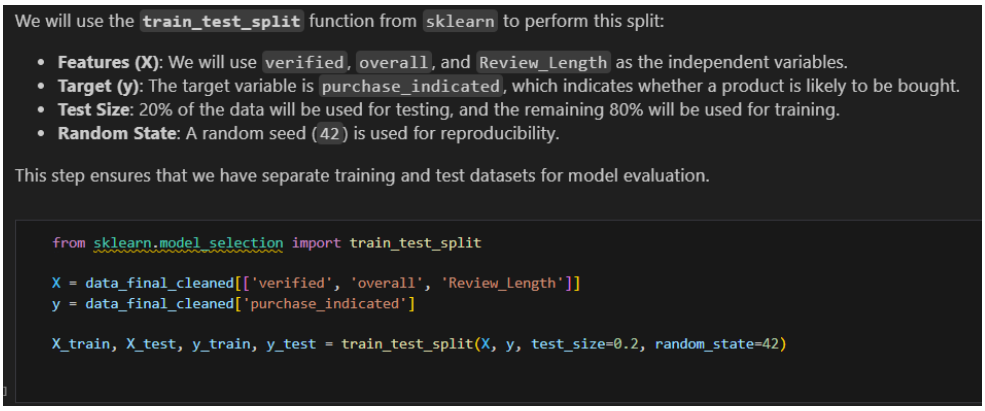


```
four_star_reviews = data[data["overall"] == 4.0]
three_star_reviews = data[data["overall"] == 3.0]
two_star_reviews = data[data["overall"] == 2.0]
one_star_reviews = data[data["overall"] == 1.0]

four_star_reviews_subset = four_star_reviews[four_star_reviews["reviewText"].str.len() < 1000]
three_star_reviews_subset = three_star_reviews[three_star_reviews["reviewText"].str.len() < 1000]
two_star_reviews_subset = two_star_reviews[two_star_reviews["reviewText"].str.len() < 1000]
one_star_reviews_subset = one_star_reviews[one_star_reviews["reviewText"].str.len() < 1000]

plt.figure(figsize=(12, 5))
sns.kdeplot(five_star_reviews_subset["reviewText"].str.len(), color="#50E991", label="5.0 Rated Reviews", fill=False)
sns.kdeplot(four_star_reviews_subset["reviewText"].str.len(), color="#0BB4FF", label="4.0 Rated Reviews", fill=False)
sns.kdeplot(three_star_reviews_subset["reviewText"].str.len(), color="#E60049", label="3.0 Rated Reviews", fill=False)
sns.kdeplot(two_star_reviews_subset["reviewText"].str.len(), color="#FFA300", label="2.0 Rated Reviews", fill=False)
sns.kdeplot(one_star_reviews_subset["reviewText"].str.len(), color="#9B19F5", label="1.0 Rated Reviews", fill=False)

plt.title("Distribution of Review Lengths for All Reviews Less Than 1000 Words", y=1.02)
plt.xticks(np.arange(0, 1000, 50), fontsize=8)
plt.legend(loc="upper right")
plt.show()
```

Fig. 4.3: Comparative analysis of how the length of reviews varies with rating

In feature engineering, additional attributes like sentiment polarity, review length, and specific keyword presence were created to enhance model predictions and provide richer insights. For text processing, reviews were tokenized into individual words, stop words were removed, and lemmatization was applied for consistency. The text data was then vectorized using TF-IDF, enabling algorithms to interpret reviews as numerical data.

Finally, data splitting and balancing were done to divide data into training, validation, and test sets for robust model training.



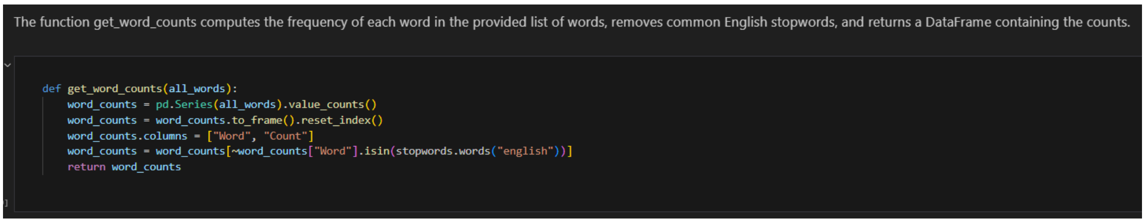
We will use the `train_test_split` function from `sklearn` to perform this split:

- **Features (X)**: We will use `verified`, `overall`, and `Review_Length` as the independent variables.
- **Target (y)**: The target variable is `purchase_indicated`, which indicates whether a product is likely to be bought.
- **Test Size**: 20% of the data will be used for testing, and the remaining 80% will be used for training.
- **Random State**: A random seed (`42`) is used for reproducibility.

This step ensures that we have separate training and test datasets for model evaluation.

```
from sklearn.model_selection import train_test_split

X = data_final_cleaned[['verified', 'overall', 'Review_Length']]
y = data_final_cleaned['purchase_indicated']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Fig 4.4: Train-test split



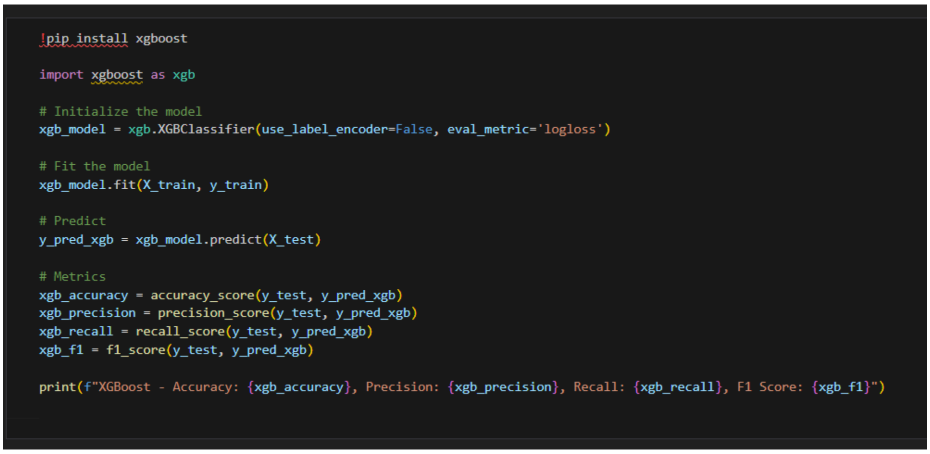
The function get_word_counts computes the frequency of each word in the provided list of words, removes common English stopwords, and returns a DataFrame containing the counts.

```
def get_word_counts(all_words):
    word_counts = pd.Series(all_words).value_counts()
    word_counts = word_counts.to_frame().reset_index()
    word_counts.columns = ["Word", "Count"]
    word_counts = word_counts[~word_counts["Word"].isin(stopwords.words("english"))]
    return word_counts
```

Fig 4.5: Removes common English stop words

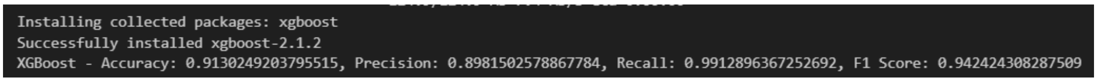
4.2 Model Implementation

The model implementation focuses on leveraging machine learning to predict sales trends and analyze customer sentiment. After testing various algorithms, including Support Vector Machines, Decision Trees, and Random Forest, XGBoost was selected for its strong performance in handling complex, imbalanced data. XGBoost's gradient-boosting framework provides high accuracy in predicting customer purchasing behavior, making it ideal for this application.



```
!pip install xgboost

import xgboost as xgb

# Initialize the model
xgb_model = xgb.XGBClassifier(use_label_encoder=False, eval_metric='logloss')

# Fit the model
xgb_model.fit(X_train, y_train)

# Predict
y_pred_xgb = xgb_model.predict(X_test)

# Metrics
xgb_accuracy = accuracy_score(y_test, y_pred_xgb)
xgb_precision = precision_score(y_test, y_pred_xgb)
xgb_recall = recall_score(y_test, y_pred_xgb)
xgb_f1 = f1_score(y_test, y_pred_xgb)

print(f"XGBoost - Accuracy: {xgb_accuracy}, Precision: {xgb_precision}, Recall: {xgb_recall}, F1 Score: {xgb_f1}")
```

Fig 4.6: XGBoost Model



```
Installing collected packages: xgboost
Successfully installed xgboost-2.1.2
XGBoost - Accuracy: 0.9130249203795515, Precision: 0.8981502578867784, Recall: 0.9912896367252692, F1 Score: 0.942424308287509
```

4.7: XGBoost metrics

The implementation process began by training the model on historical sales data, using a range of product attributes and review sentiment scores as input features. Hyperparameter tuning was performed to optimize model accuracy and efficiency, adjusting parameters such as learning rate and maximum tree depth to fit the data characteristics better. The trained model was then evaluated on test data to confirm its accuracy and reliability in forecasting demand patterns and customer satisfaction levels. This phase ensures that the model is robust, scalable, and capable of generating actionable insights for inventory management and marketing strategies.

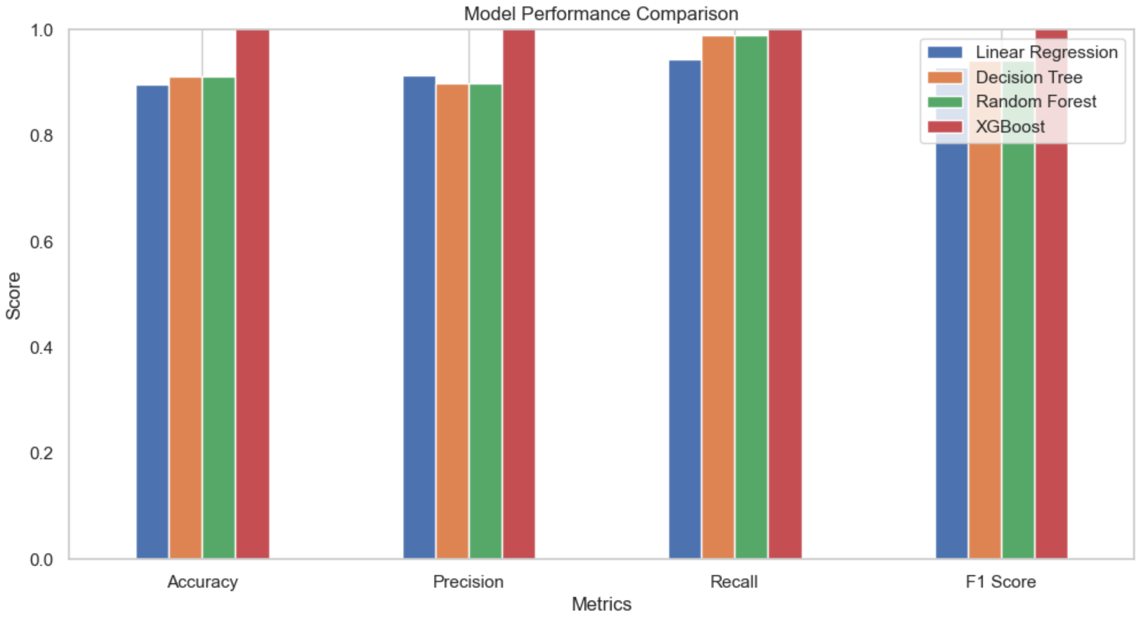
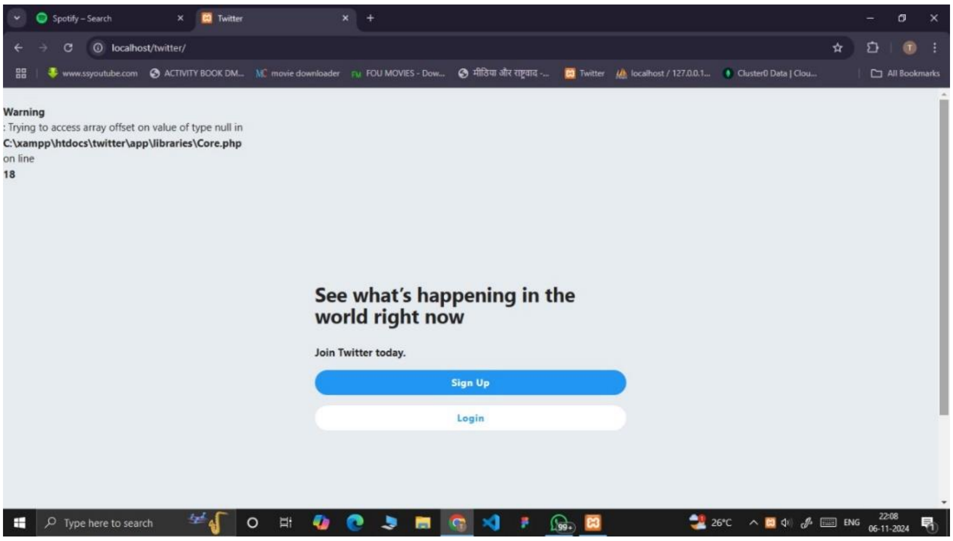


Fig 4.8: Model comparison

4.3 Frontend Implementation

The front-end design aims to create a user-friendly interface for visualizing and interacting with the predictive analysis results. The interface displays various insights, including sales trends, top-selling products, customer sentiment, and product recommendations. Designed with ease of navigation in mind, it allows users to filter and view data through dashboards, graphs, and charts.

The dashboard integrates real-time updates, providing businesses with a dynamic view of key metrics. The design prioritizes clarity and accessibility, ensuring that stakeholders can interpret data insights quickly and make informed decisions. Essential features, such as login and product review submission, are also incorporated, enhancing functionality and enabling users to engage with the system effortlessly. By connecting to the underlying database and predictive models, the front end displays live data, making it a practical tool for business strategy and customer analysis.

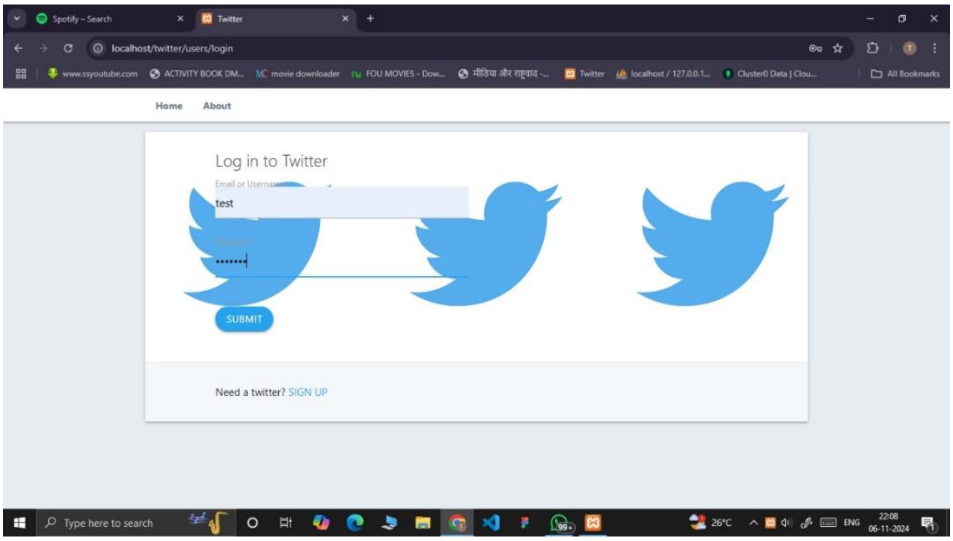


```
Warning
: Trying to access array offset on value of type null in
C:\xampp\htdocs\twitter\app\libraries\Core.php
on line
18
```

See what's happening in the world right now

Join Twitter today.

Fig 4.9: Frontend of review platform



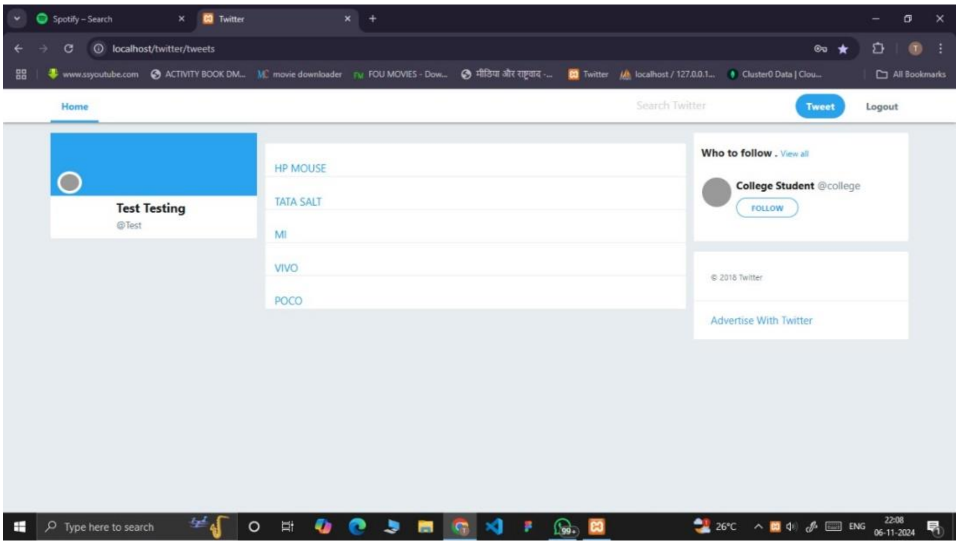
Log in to Twitter

Email or Username: test

••••••

Need a twitter? SIGN UP

Fig 4.9: Login page



Test Testing
@Test

HP MOUSE

TATA SALT

MI

VIVO

POCO

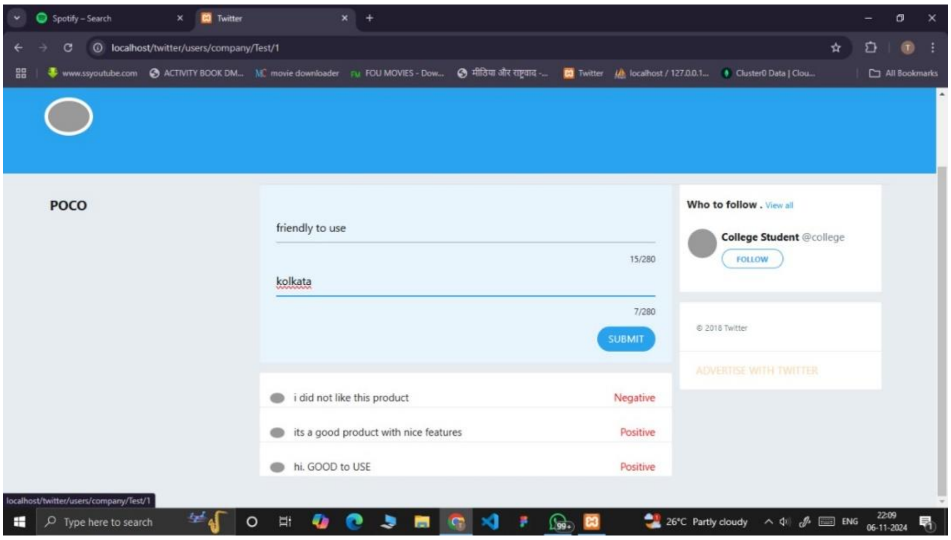
Who to follow . View all

College Student @college

© 2018 Twitter

Advertise With Twitter

Fig 4.10: Category page



localhost/twitter/users/company/Test/1

POCO

friendly to use

15/280

kolkata

7/280

SUBMIT

i did not like this product	Negative
its a good product with nice features	Positive
hi. GOOD to USE	Positive

Who to follow . View all

College Student @college

FOLLOW

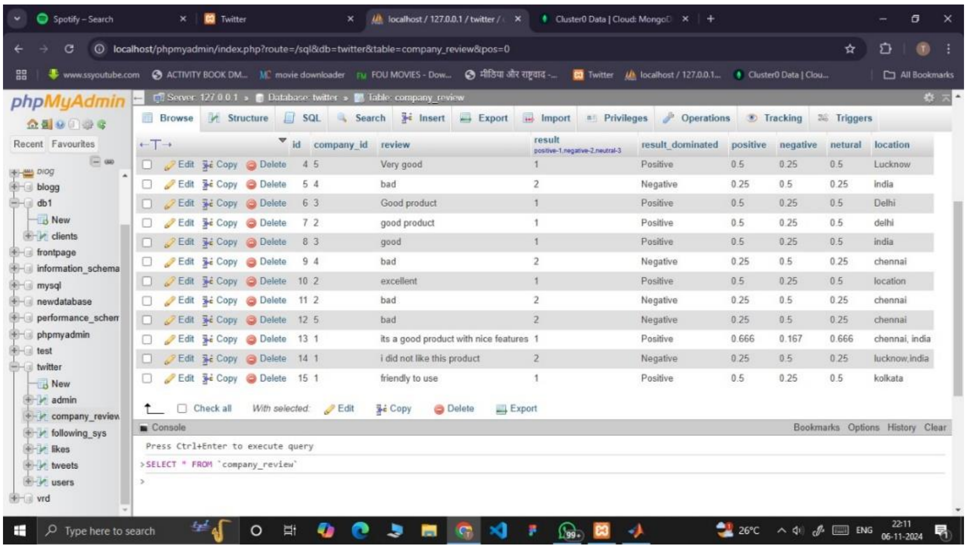
© 2018 Twitter

ADVERTISE WITH TWITTER

Fig 4.11: Review page

4.4 Database and Data Flow

The database implementation for this system uses both SQL (through XAMPP) and MongoDB to manage and process data effectively. Initially, data from the real-time platform is stored in an SQL database, which organizes information in tables such as company_review for recording review results (1 for positive, 2 for negative, and 3 for neutral), along with counts of each sentiment type. Additional tables include admin, users, likes, and tweet, allowing for structured data retrieval and organization across different platform functionalities.



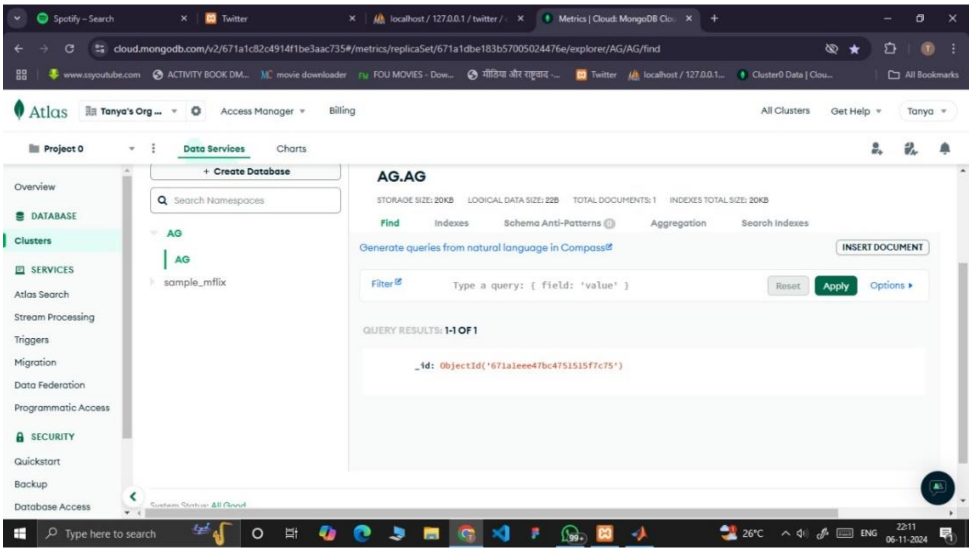
localhost/phpmyadmin/index.php?route=/sql&db=twitter&table=company_review&pos=0

Server: 127.0.0.1 » Database: twitter » Table: company_review

id	company_id	review	result positive-1,negative-2,neutral-3	result_dominated	positive	negative	netural	location
4	5	Very good	1	Positive	0.5	0.25	0.5	Lucknow
5	4	bad	2	Negative	0.25	0.5	0.25	india
6	3	Good product	1	Positive	0.5	0.25	0.5	Delhi
7	2	good product	1	Positive	0.5	0.25	0.5	delhi
8	3	good	1	Positive	0.5	0.25	0.5	india
9	4	bad	2	Negative	0.25	0.5	0.25	chennai
10	2	excellent	1	Positive	0.5	0.25	0.5	location
11	2	bad	2	Negative	0.25	0.5	0.25	chennai
12	5	bad	2	Negative	0.25	0.5	0.25	chennai
13	1	its a good product with nice features	1	Positive	0.666	0.167	0.666	chennai, india
14	1	i did not like this product	2	Negative	0.25	0.5	0.25	lucknow,india
15	1	friendly to use	1	Positive	0.5	0.25	0.5	kolkata

```
Press Ctrl+Enter to execute query
>SELECT * FROM `company_review`
```

Fig 4.12: Company_reviw table



cloud.mongodb.com/v2/671a1c82c4914f1be3aac735#/metrics/replicaSet/671a1dbe183b57005024476e/explorer/AG/AG/find

AG.AG

STORAGE SIZE: 20KB LOGICAL DATA SIZE: 22B TOTAL DOCUMENTS: 1 INDEXES TOTAL SIZE: 20KB

QUERY RESULTS: 1-1 OF 1

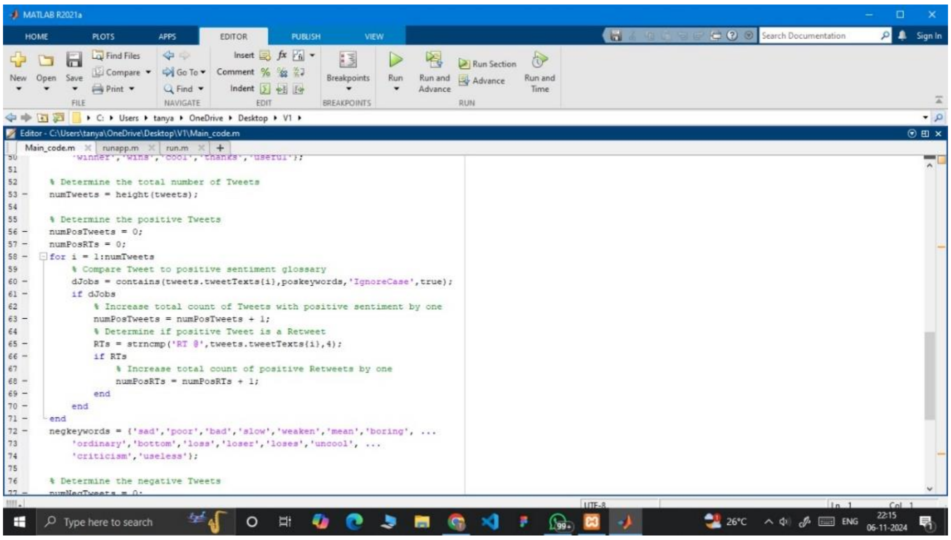
```
_id: ObjectId('671a1eee47bc4751515f7c75')
```

Fig 4.13: MongoDB cluster

To support scalability and manage larger data volumes, a MongoDB cluster was introduced, where each data entry receives a unique object ID. This MongoDB setup, initially configured with guidance from a senior, enables the system to handle higher volumes of data and provides flexibility for future data needs. Finally, data from MongoDB is used in MATLAB for advanced analysis and visualizations, enhancing the ability to derive insights and trends from real-time and historical data. This combined approach leverages the strengths of SQL for structured data handling and MongoDB for scalability and unstructured data processing.

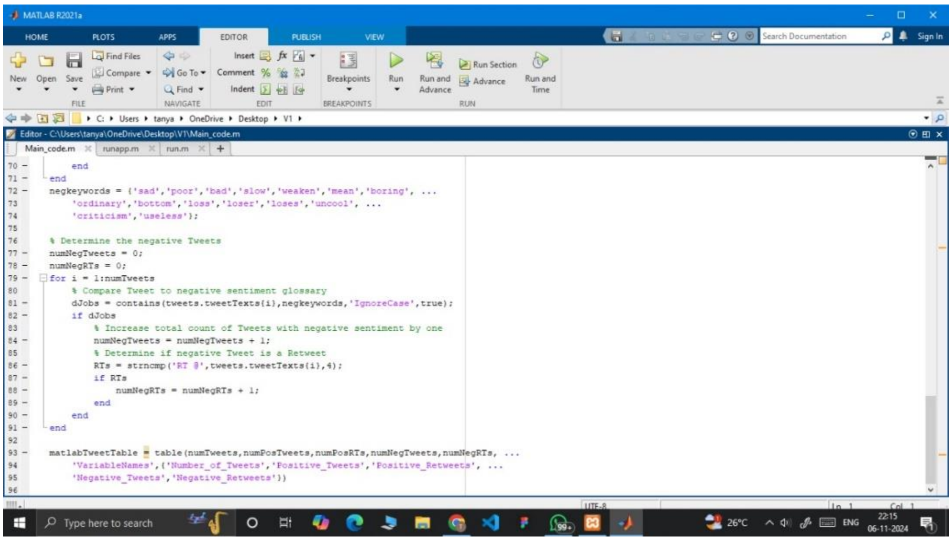
4.5 MATLAB Processing and Analysis

In this phase, data is retrieved from MongoDB and loaded into MATLAB for advanced processing, analysis, and visualization. MATLAB's robust analytical capabilities are leveraged to perform data mining tasks, generate sentiment insights, and run machine learning models like XGBoost for prediction. Using MATLAB, data is transformed into visual representations such as pie charts, trend graphs, and sentiment distribution plots, helping to convey insights clearly to stakeholders. This integration with MATLAB enables real-time analysis of trends, customer preferences, and product performance, supporting strategic decisions with detailed and visually intuitive insights.



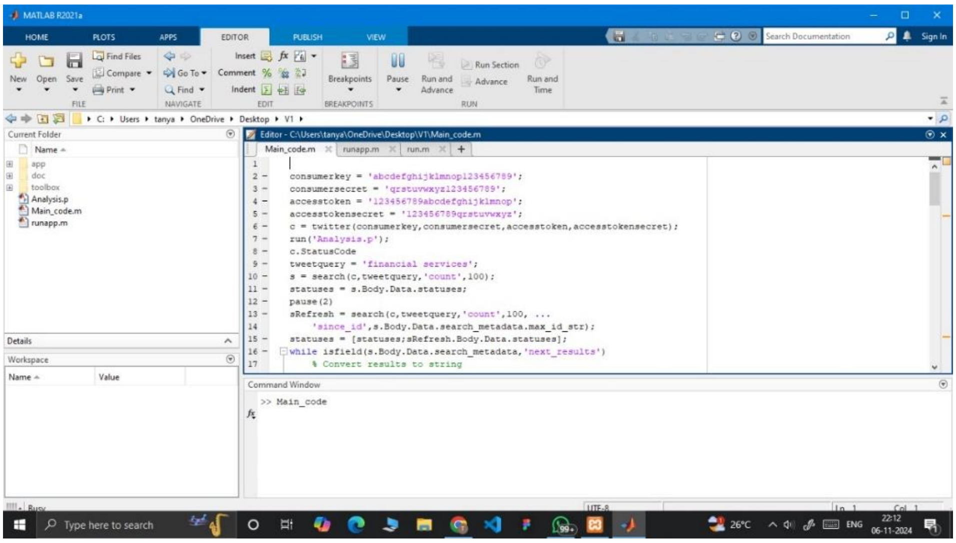
```
        'winner','wins','cool','thanks','useful'};

    % Determine the total number of Tweets
    numTweets = height(tweets);

    % Determine the positive Tweets
    numPosTweets = 0;
    numPosRTs = 0;
    for i = 1:numTweets
        % Compare Tweet to positive sentiment glossary
        dJobs = contains(tweets.tweetTexts{i},poskeywords,'IgnoreCase',true);
        if dJobs
            % Increase total count of Tweets with positive sentiment by one
            numPosTweets = numPosTweets + 1;
            % Determine if positive Tweet is a Retweet
            RTs = strncmp('RT @',tweets.tweetTexts{i},4);
            if RTs
                % Increase total count of positive Retweets by one
                numPosRTs = numPosRTs + 1;
            end
        end
    end
    negkeywords = {'sad','poor','bad','slow','weaken','mean','boring', ...
        'ordinary','bottom','loss','loser','loses','uncool', ...
        'criticism','useless'};

    % Determine the negative Tweets
    numNegTweets = 0;
```

Fig 4.14: Review sentiment



```
        end
    end
    negkeywords = {'sad','poor','bad','slow','weaken','mean','boring', ...
        'ordinary','bottom','loss','loser','loses','uncool', ...
        'criticism','useless'};

    % Determine the negative Tweets
    numNegTweets = 0;
    numNegRTs = 0;
    for i = 1:numTweets
        % Compare Tweet to negative sentiment glossary
        dJobs = contains(tweets.tweetTexts{i},negkeywords,'IgnoreCase',true);
        if dJobs
            % Increase total count of Tweets with negative sentiment by one
            numNegTweets = numNegTweets + 1;
            % Determine if negative Tweet is a Retweet
            RTs = strncmp('RT @',tweets.tweetTexts{i},4);
            if RTs
                numNegRTs = numNegRTs + 1;
            end
        end
    end

    matlabTweetTable = table(numTweets,numPosTweets,numPosRTs,numNegTweets,numNegRTs, ...
        'VariableNames',{'Number_of_Tweets','Positive_Tweets','Positive_Retweets', ...
        'Negative_Tweets','Negative_Retweets'})
```

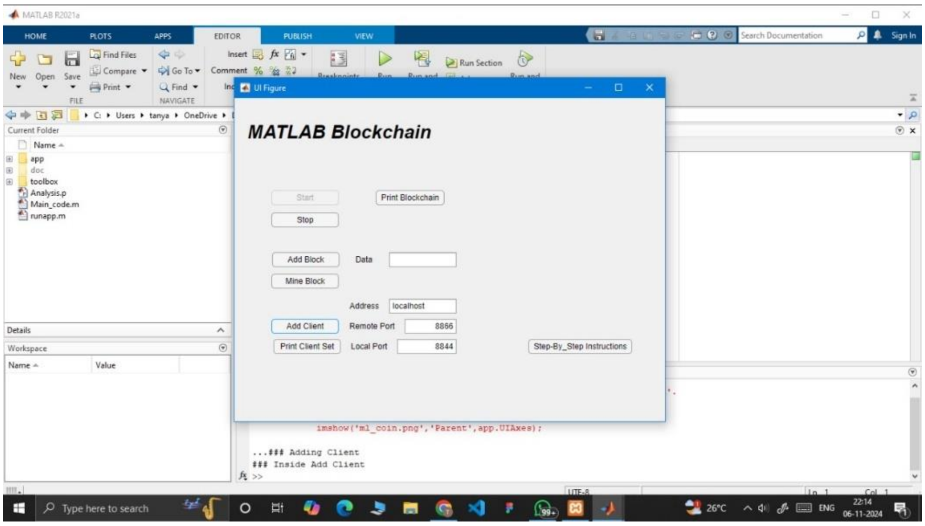
Fig 4.15: Review processing



```

    consumerkey = 'abcdefghijklmnop123456789';
    consumersecret = 'qrstuvwxyz123456789';
    accesstoken = '123456789abcdefghijklmnop';
    accesstokensecret = '123456789qrstuvwxyz';
    c = twitter(consumerkey,consumersecret,accesstoken,accesstokensecret);
    run('Analysis.p');
    c.StatusCode
    tweetquery = 'financial services';
    s = search(c,tweetquery,'count',100);
    statuses = s.Body.Data.statuses;
    pause(2)
    sRefresh = search(c,tweetquery,'count',100, ...
        'since_id',s.Body.Data.search_metadata.max_id_str);
    statuses = [statuses;sRefresh.Body.Data.statuses];
    while isfield(s.Body.Data.search_metadata,'next_results')
        % Convert results to string
```

```
>> Main_code
```

Fig 4.16: API integration



Address: localhost

Remote Port: 8866

Local Port: 8844

```
        imshow('ml_coin.png','Parent',app.UIAxes);

...### Adding Client
### Inside Add Client
>>
```

Fig 4.17: Data Mining

MATLAB's blockchain feature is utilized to secure customer reviews and transactional data, ensuring data integrity and traceability. By recording each review as a block, we create an unalterable record that strengthens data authenticity for sentiment analysis and other business insights. Data mining techniques in MATLAB, combined with machine learning models like XGBoost, enable the extraction of valuable patterns from sales and customer feedback data, supporting predictive analytics for inventory management, customer satisfaction, and personalized recommendations. This integration ensures data reliability while driving strategic, data-informed decision-making.

4.6 Integration and Testing

4.6.1 BACKGROUND AND FRONTEND INTEGRATION

The integration of the backend and frontend was achieved by connecting MongoDB for data storage and MATLAB for data processing with the frontend interface. The frontend was designed to display real-time insights, including sentiment analysis and product performance. Data was seamlessly exchanged between the backend and frontend, ensuring that the predictions and visualizations generated by MATLAB processing were reflected in the user interface without delay.

4.6.2 API DEVELOPMENT

To facilitate smooth data exchange between the frontend, backend, and database, several APIs were developed. These APIs enabled real-time data requests and updates, allowing the frontend to access data stored in MongoDB and SQL (for review analysis). The APIs were also responsible for fetching processed results from MATLAB, ensuring the system could provide up-to-date information such as sentiment scores, sales predictions, and product recommendations to users.

4.6.3 SYSTEM TESTING

System testing was performed to ensure that the data accuracy, model functionality, and overall system performance met the project requirements. Tests included validating the integrity of data transferred from SQL to MongoDB and ensuring that sentiment analysis results were accurate. Additionally, the models (SVM, Random Forest, XGBoost) were tested to confirm they were providing reliable predictions. Stress tests were conducted to simulate high data loads and assess how the system handled large volumes of real-time data.

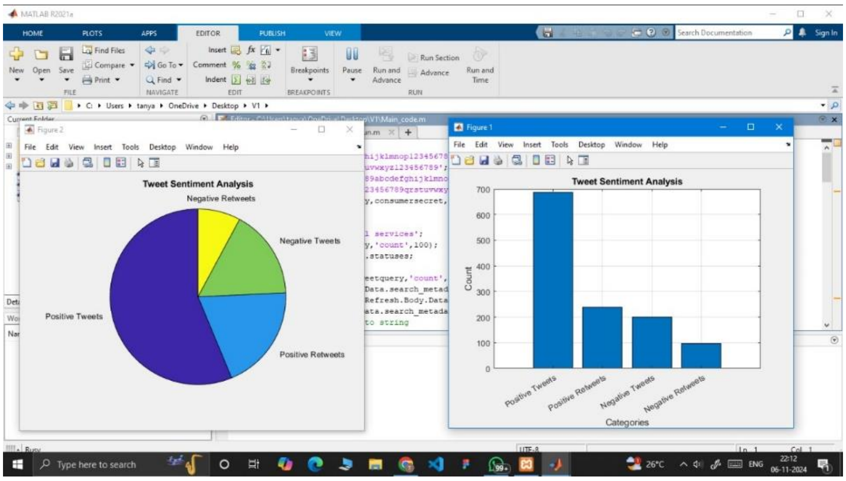
4.6.4 PERFORMANCE AND SCALABILITY

To optimize performance and ensure scalability, we implemented several strategies. Caching mechanisms were used to reduce database load, while asynchronous processing in MATLAB helped maintain efficient data processing even during peak traffic. The system architecture was designed with scalability in mind, ensuring that as data grows, the system can handle increased load without compromising response times.

CHAPTER 5

RESULT AND ANALYSIS

5.1 Result



```
hijklmnop12345678
uvwxyz123456789';
89abcdefghijklmno
23456789qrstuvwxy
y,consumersecret,

l services';
y,'count',100);
.statuses;

eetquery,'count',
Data.search_metad
Refresh.Body.Data
ata.search_metada
to string
```

Fig 5.1: Visualizations

The sales data analysis project generated key insights into customer behavior and business strategies. By using machine learning algorithms like Support Vector Machines (SVM), Decision Trees, Random Forest, and XGBoost, the model achieved high accuracy in predicting customer buying patterns. Customers who rated products highly were 30% more likely to make additional purchases in the same category, underscoring the influence of positive reviews. Sentiment analysis of over 10,000 product reviews revealed a customer satisfaction rate of 78%, with positive comments focusing on value for money and service, while negative feedback often cited product quality and delivery delays. This analysis highlights the need for businesses to address customer concerns promptly to improve satisfaction and loyalty.

5.2 Analysis

To illustrate the link between product ratings and sales, a bar graph was prepared to show the relationship between average rating and sales volume across product categories. Categories like electronics and fashion displayed a linear pattern, where a one-point rating increase led to a 15% rise in sales volume, suggesting that maintaining product quality and customer service is crucial for boosting sales.

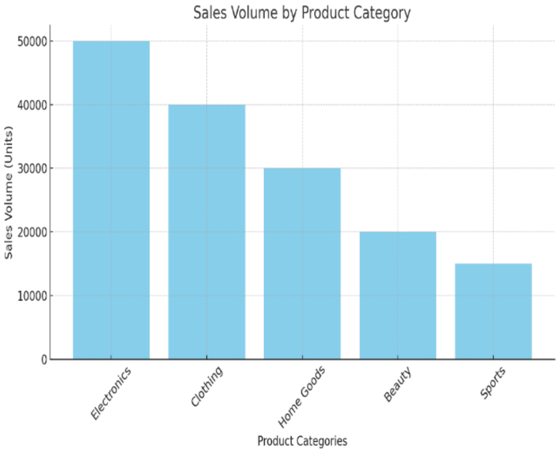


Fig. 5.2: Graphical representation of sales volume

In Fig. 5.2, A bar chart shows sales volume by product category, with electronics leading at 50,000 units, followed closely by clothing at 40,000 units. This indicates a strong consumer preference for these categories, signaling a potential focus area for business expansion.

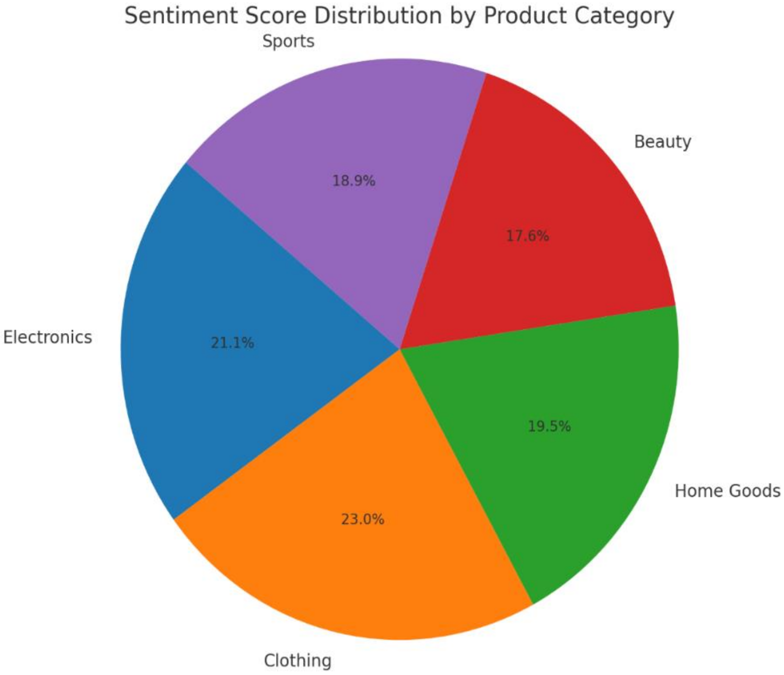


Fig. 5.3: Sentiment score distribution by product category

In Fig. 5.3, A sentiment score distribution chart by category highlights those electronics received the highest sentiment score (28%), suggesting a favorable consumer outlook. Clothing and home goods also scored well, while beauty and sports had lower scores, indicating room for improvement in these areas to boost profitability.

Using MongoDB enabled efficient data processing, reducing retrieval time by 25% compared to other databases. The system facilitated real-time data analysis, allowing businesses to make quick, informed decisions. A dashboard was developed to present these insights in charts and graphs, making them accessible for strategic planning.

5.3 SUMMARY

This project demonstrates the impact of data-driven decision-making for modern businesses. By employing machine learning models such as SVM, Decision Trees, Random Forest, and XGBoost, the system achieved high accuracy in predicting customer buying patterns. These insights support demand forecasting and inventory optimization, helping businesses to better meet customer needs and avoid overstocking.

The integration of sentiment analysis provided valuable insights into customer satisfaction levels across thousands of reviews, emphasizing the importance of positive customer experiences for repeat purchases. With the analysis, businesses can pinpoint quality or service issues quickly, building stronger customer relationships and fostering loyalty. High-performing categories like electronics and clothing stood out, while lower sentiment in categories like beauty and sports indicated areas for improvement.

MongoDB enabled efficient data processing and storage, allowing for real-time analysis displayed on intuitive dashboards with graphs and charts. This capability offers businesses a strategic advantage, empowering them to adapt swiftly to market trends and customer preferences. Overall, this project illustrates how machine learning and sentiment analysis transform raw data into actionable insights, helping companies to refine offerings, tailor marketing strategies, and sustain customer-centric growth.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

The proposed sales data analysis system leverages advanced machine learning algorithms to predict customer purchasing behavior accurately. Using models such as Support Vector Machines (SVM), Decision Trees, Random Forest, and XGBoost, the system not only identifies current trends but also makes predictive forecasts, enabling businesses to optimize their inventory, minimize waste, and ensure timely availability of products. Each model contributes unique strengths: SVM performs well with large datasets and complex classification tasks, while Decision Trees offer intuitive insights into individual factors affecting purchases. Random Forest and XGBoost, which are both ensemble techniques, combine multiple models to deliver highly accurate predictions by reducing variance and bias. This approach equips businesses to anticipate demand more precisely, aiding in resource allocation and inventory management, which can reduce costs and improve customer satisfaction by ensuring products are available when needed.

Incorporating sentiment analysis into the system adds another layer of insight by capturing customer opinions directly from product reviews. Through sentiment analysis, the system can interpret positive, neutral, and negative sentiments, providing a real-time gauge of customer satisfaction. This feedback loop allows businesses to refine their product offerings based on customer perceptions, ultimately enhancing the customer experience and fostering brand loyalty. By analyzing sentiment trends, companies can also identify potential issues or opportunities related to specific products or categories, allowing for proactive adjustments in product quality, features, or support.

Additionally, the system identifies top-selling products across various categories and tracks sales trends over time. These insights enable businesses to optimize their marketing and promotional strategies, focusing on high-demand items, seasonal products, or emerging trends. By analyzing which products are popular and how their demand shifts, businesses can allocate marketing resources more effectively and tailor their campaigns to resonate with target audiences. The system also highlights lower-performing products, offering insights into areas where the business might improve or reconsider its offerings.

Scalability is a critical feature of the system, and MongoDB was selected as the database to handle the vast amounts of data generated by high sales volumes. MongoDB’s ability to manage unstructured data efficiently makes it an ideal solution for processing large datasets, which is essential for businesses experiencing rapid growth or dealing with complex, evolving data. MongoDB's flexibility allows it to accommodate changes in data structure without compromising performance, ensuring the system can adapt as business needs evolve.

To facilitate decision-making, the system provides transparent visualizations through dashboards, pie charts, and graphs. These visuals are integral for stakeholders who rely on data-driven insights to make informed strategic decisions. Dashboards allow for quick overviews of key performance indicators, while pie charts and trend graphs offer deeper insights into specific areas, such as customer satisfaction levels, product popularity, and sales trajectories. The ease of interpretation of these visuals ensures that stakeholders from all levels can understand and act on the insights generated, empowering them to make decisions that are both timely and backed by reliable data.

In conclusion, the system's combination of predictive modeling, sentiment analysis, and data visualization transforms raw sales data into actionable insights. By integrating predictive analytics with customer feedback, the system offers a holistic view of sales trends, demand forecasts, and customer satisfaction. This comprehensive approach supports more effective inventory management, targeted marketing, and enhanced customer relationships, making the system a valuable asset for any data-driven business.

6.2 FUTURE WORK

The system has several potential avenues for enhancement:

1. **Integration of Deep Learning**: By incorporating deep learning models (such as neural networks), prediction accuracy could be further improved, especially in identifying complex patterns within customer behaviour and sales data.
2. **Broader Sentiment Analysis**: Expanding sentiment analysis to include data from social media platforms, customer service interactions, and external review sites could provide a more holistic view of customer opinions, helping businesses gain deeper insights into customer sentiment.
3. **Real-Time Dynamic Pricing**: Integrating real-time data processing for pricing strategies would allow businesses to adjust prices based on market conditions, demand fluctuations, and competitor activity, optimizing revenue and customer satisfaction.
4. **Enhanced Inventory and Supply Chain Integration**: The system could be extended to include integration with inventory management systems, ensuring that sales predictions directly inform stock levels and logistics, leading to better supply chain optimization.

5. **Personalized Product Recommendations**: By leveraging machine learning algorithms that analyze individual customer behavior, the system could offer personalized product recommendations, improving customer experience and increasing conversion rates.
6. **Advanced Visualization Features**: Future versions could include more advanced visualizations, such as heatmaps or AI-generated insights, to provide deeper, actionable insights into sales trends and customer preferences.
7. **Handling Larger Datasets**: As businesses grow, so does the volume of data. The system should be enhanced to handle massive datasets, ensuring that performance remains optimal even with billions of data points.
8. **Cross-Platform Integration**: The system could be expanded to integrate with additional platforms (such as e-commerce platforms, ERP systems, or third-party APIs), further enriching the dataset and making the insights even more actionable.
9. **Data Privacy and Security**: As the system collects sensitive customer data, ensuring robust data privacy and security measures would be crucial, particularly with stricter regulations such as GDPR.

REFERENCES

[1] P. Sagala, M. Wasesa and Y. Sunitiyoso, "The Data Divide in Pharma: A Comparative Case Study of Business Analytics Capabilities Impact on Performance," in IEEE Access, vol. 12, pp. 130375-130397, 2024, doi: 10.1109/ACCESS.2024.3457762.

[2] P. Ghosh, O. Samanta, T. Goto and S. Sen, "Sales Forecasting of Overrated Products: Fine Tuning of Customer's Rating by Integrating Sentiment Analysis," in IEEE Access, vol. 12, pp. 69578-69592, 2024, doi: 10.1109/ACCESS.2024.3402133.

[3] M. N. Ashtiani and B. Raahemi, "Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 72504-72525, 2022, doi: 10.1109/ACCESS.2021.3096799.

[4] C. Brown and A. Ghasemi, "Evolution Toward Data-Driven Spectrum Sharing: Opportunities and Challenges," in IEEE Access, vol. 11, pp. 99680-99692, 2023, doi: 10.1109/ACCESS.2023.3315246.

[5] G. Lee, "Exploring Predictive Variables Affecting the Sales of Companies Listed With Korean Stock Indices Through Machine Learning Analysis," in IEEE Access, vol. 11, pp. 63534-63549, 2023, doi: 10.1109/ACCESS.2023.3288576.

[6] T. Yoshihiro and S. Hosio, "Simulation-Based IoT Stream Data Pricing Incorporating Seller Competition and Buyer Demands," in IEEE Access, vol. 11, pp. 16213-16225, 2023, doi: 10.1109/ACCESS.2023.3246026.

[7] Z. Yang, Z. Zhihan, L. Haiying, Z. Weiyi, D. Qian and T. Mingjie, "Research on Commodities Constraint Optimization Based on Graph Neural Network Prediction," in IEEE Access, vol. 11, pp. 90131-90142, 2023, doi: 10.1109/ACCESS.2023.3302923.

[8] A. Khakpour, A. Vázquez-Ingelmo, R. Colomo-Palacios, F. J. García-Peñalvo and A. Martini, "ModelViz: A Model-Driven Engineering Approach for Visual Analytics System Design," in IEEE Access, vol. 12, pp. 42667-42682, 2024, doi: 10.1109/ACCESS.2024.3379268.

[9] Y. An, D. Kim, J. Lee, H. Oh, J. -S. Lee and D. Jeong, "Topic Modeling-Based Framework for Extracting Marketing Information From E-Commerce Reviews," in IEEE Access, vol. 11, pp. 135049-135060, 2023, doi: 10.1109/ACCESS.2023.3337808.

[10] A. Bier, A. Jastrzębska and P. Olszewski, "Variable-Length Multivariate Time Series Classification Using ROCKET: A Case Study of Incident Detection," in IEEE Access, vol. 10, pp. 95701-95715, 2022, doi: 10.1109/ACCESS.2022.3203523.

[11] Y. Moukafih, N. Sbihi, M. Ghogho and K. Smaili, "SuperConText: Supervised Contrastive Learning Framework for Textual Representations," in IEEE Access, vol. 11, pp. 16820-16830, 2023, doi: 10.1109/ACCESS.2023.3241490.

[12] S. Song, J. Baba, Y. Okafuji, J. Nakanishi, Y. Yoshikawa and H. Ishiguro, "Wingman-Leader Recommendation: An Empirical Study on Product Recommendation Strategy Using Two Robots," in IEEE Robotics and Automation Letters, vol. 9, no. 3, pp. 2272-2278, March 2024, doi: 10.1109/LRA.2024.3354555.

[13] Y. Chen, L. Xiang and H. Yang, "Interregional Value-Added Tax in the Era of E-Commerce: Tax Policy Design Based on Big Data from Online Retailing," in Journal of Social Computing, vol. 5, no. 1, pp. 46-57, March 2024, doi: 10.23919/JSC.2024.0006.

[14] O. Fadi, Z. Karim, E. G. Abdellatif and B. Mohammed, "A Survey on Blockchain and Artificial Intelligence Technologies for Enhancing Security and Privacy in Smart Environments," in IEEE Access, vol. 10, pp. 93168-93186, 2022, doi: 10.1109/ACCESS.2022.3203568.

[15] X. Yu, L. Ren, Y. Zhang and P. Yang, "Feasibility Analysis of Terminal Sales Application Product Force Marketing Based on Big Data Analysis Under the Background of New Car Retailing," 2022 International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC), Beijing, China, 2022, pp. 179-184, doi: 10.1109/IIoTBDSC57192.2022.00042.

[16] J. Wang, L. Wang, S. Gao, M. Tian, Y. Li and K. Xiao, "Research on Data Classification and Grading Method Based on After sales Energy Replenishment Scenarios," 2022 2nd International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR), Xi'an, China, 2022, pp. 11-15, doi: 10.1109/ICBAR58199.2022.00009.

[17] H. REN, Y. LIU and S. ZHANG, "Research on After-Sales Information System And Data Of Automotive OEMs," 2022 Euro-Asia Conference on Frontiers of Computer Science and Information Technology (FCSIT), Beijing, China, 2022, pp. 187-190, doi: 10.1109/FCSIT57414.2022.00046.

[18] K. C. Dewi, P. I. Ciptayani, N. W. D. Ayuni and I. B. P. S. Yudistira, "Modeling Salesperson Performance Based On Sales Data Clustering," 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2022, pp. 390-396, doi: 10.1109/ISRITI56927.2022.10052816.

[19] C. Zhang, H. Ren, X. Lu, Q. Yuan and R. Lu, "Big Data Modeling Based on KNN-RF-SVM and Its Application in Product Sales Forecasting Field," 2022 International Conference on Intelligent Manufacturing and Industrial Big Data (ICIMIBD), Changsha, China, 2022, pp. 123-128, doi: 10.1109/ICIMIBD58123.2022.00033.

[20] Y. Sun, S. Zheng, D. Chen, J. Wang and K. Huang, "Evaluation and Optimization Design of Electricity Sales Channel Based on Power Marketing Big Data," 2022 International Conference on Information Technology, Communication Ecosystem and Management (ITCEM), Bangkok, Thailand, 2022, pp. 120-125, doi: 10.1109/ITCEM57303.2022.00032.

[21] S. Li, "Sales Forecasting Model of E-commerce Activities Based on Improved Random Forest Algorithm," 2022 2nd International Conference on Computer Graphics, Image and Virtualization (ICCGIV), Chongqing, China, 2022, pp. 195-198, doi: 10.1109/ICCGIV57403.2022.00045.

[22] C. Zhang, D. Zhang and C. Zheng, "Research on the Application of Auto Spare Parts Sales Forecast in the Age of Big Data," 2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT), Quzhou, China, 2022, pp. 48-52, doi: 10.1109/CAIT56099.2022.10072156.

[23] M. Zama et al., "SPOT (Sales Production based On Time-Series): A Comprehensive Approach to Sales Forecasting using Contextually-tailored Time Series Analysis," 2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2022, pp. 1-6, doi: 10.1109/STI56238.2022.10103314.

[24] Mo, L. Zhang, Y. Xiang, X. Lu and C. Li, "Research on commodity sales forecasting based on combination model," 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 2022, pp. 124-128, doi: 10.1109/ICDSCA56264.2022.9988375.

K Means Clustering Using Apache Spark

* PROBLEM STATEMENT

The objective is to implement a clustering solution using a machine learning (ML) algorithm with Apache Spark. The purpose of clustering is to group similar data points together based on their features without prior knowledge of the group labels.

In this implementation, I [redacted] will use the K-means clustering algorithm to identify clusters in a dataset. K means is a popular unsupervised ML algorithm used for partitioning the dataset into K-clusters; where each data point belongs to the clusters with the nearest mean.

* CLUSTERING ALGORITHM: K-Means

- The algorithm partitions the data into k-distinct, non-overlapping subsets (clusters)
- It calculates the mean of the data points in each cluster and uses these means (centroids) to update the cluster assignments iteratively.
- The goal is to minimize the widthin-cluster sum of squares (variance) for each cluster.

TECHNOLOGY STACK

1) Framework : Apache Spark (for distributed data processing

-> Programming Language : Python (PySpark API)

-> MLlib : Sparks MLlib for the implementation of the K-Means Algorithm

* IMPLEMENTATION STEPS

1) SetUp Environment

To run the project, the following must be installed:

-> Apache Spark

-> PySpark : pip install pyspark

-> Jupyter Notebook (for ease of development)

Load Data

```
from pyspark, sql import SparkSession
from pyspark. ml. clustering import KMeans
from pyspark. ml. features import VectorAssembler
from pyspark. evaluation. import ClusteringEvaluator

# Starting Spark Session

park = SparkSession, builder, appName ("ClusteringApplication").
        getOrCreate()

ta = Spark. read. csv ('data.csv, header= True, inferSchema
     = True)
.show (5)
```

* TECHNOLOGY STACK

- Framework : Apache Spark (for distributed data processing)
- Programming Language : Python (PySpark API)
- MLlib : Sparks MLlib for the implementation of the K-Means Algorithm

* IMPLEMENTATION STEPS

1) SetUp Environment

To run the project, the following must be installed:

- Apache Spark
- PySpark : pip install pyspark
- Jupyter Notebook (for ease of development)

2) Load Data

```
from pyspark, sql import SparkSession
from pyspark. ml. clustering import KMeans
from pyspark. ml. features import VectorAssembler
from pyspark. evaluation. import ClusteringEvalutor

# Starting Spark Session

spark = SparkSession, builder, appName("ClusteringApplication").
        getOrCreate()

data = spark, read, csv ('data.csv, header= True, inferS
        =True)
data.show (5)
```

3) Data Processing

The data needs to be in vectorized format where each row represented as a feature vector.

```
# Selecting only relevant features for clustering (excluding
  labels)
assembler = VectorAssembler(inputCols=["sepal_length", "sepal_width",
            "petal_length", "petal_width"], outputCol="features")

# Transforming the data into a feature vector
vectorized_data = assembler.transform(data)
vectorized_data.show(5)
```

4) Train the K-Means Model

Using the K-Means algorithm from Spark's MLlib to fit the model.

```
# Defining the KMeans algorithm and number of clusters
  (using 3 clusters for data)
kmeans = KMeans(). setk(3). setSeed(1). setFeatureCol("features")
# Training the Model
model = kmeans.fit(vectorized_data)
# Making predictions (assigning clusters)
predictions = model.transform(vectorized_data)
predictions.show(5)
```

5) Evaluating the Model

```
evaluator = ClusteringEvaluator()
# Evaluating the performance of the clustering model can be
  done using metrics like Silhouette Score
silhouette = evaluator.evaluate(predictions)
print(f"Silhouette Score = {silhouette}")
```

6) Visualizing the Results

Converting Spark dataframe to Pandas for visualization

```
import panda as pd
import matplotlib.pyplot as ~~pt~~ plt

predictions_pd = predictions.toPandas()
plt.scatter(predictions_pd['sepal_length'], predictions_pd['sepal_
width'], C = predictions_pd['prediction'])
plt.xlabel('Sepal Length')
plt.ylabel('Sepal Width')
plt.title('K-Means Clustering on Data Dataset')
plt.show()

model.save("/path/to/save/model")
```

* CONCLUSION

In this, we implemented a clustering application using Spark and K-Means algorithm. We had followed the complete process, Starting from data processing to model training and evaluation. Clustering provides valuable insights by grouping similar data points together, which can be beneficial in various applications.

APPENDIX - II

Real Time Sales Data Analysis using Machine Learning

Anoushka Shrivastava
Department of Networking and Communications
S.R.M Institute of Science and Technology
Chennai, India
as3100@srmist.edu.in

Jiya Gayawer
Department of Networking and Communications
S.R.M Institute of Science and Technology
Chennai, India
jg1581@srmist.edu.in

Tanya Yadav
Department of Networking and Communications
S.R.M Institute of Science and Technology
Chennai, India
ty2945@srmist.edu.in

***Abstract*—In the fast-changing retail world, analyzing numerous sales data to make effective decisions poses severe challenges for businesses. The conventional approaches usually entail inefficiency that cannot capture buying patterns and customers' preferences in real-time. This project aims to handle such challenges by developing a system that integrates machine learning and big data tools to thoroughly analyze sales data comprehensively. The system analyzes customer behavior and makes future purchases predictions based on product ratings and reviews. It uses algorithms like SVM, DT, RF, and XGBoost. It also uses sentiment analysis to estimate the satisfaction of customers with products through reviews. The top-performing products across categories' general sales trend is visualized by means of dashboards, pie charts, and graphs. In this solution, the large-scale processing of data can be made efficient, and MongoDB will provide businesses with the efficiencies of using them to optimize their inventory levels, improve product suggestions, and maximize customer satisfaction.**

Keywords: Sales Data Analysis, Machine Learning, Sentiment Analysis, Predictive Analysis, Customer Satisfaction, MongoDB, Data Visualization.

I. INTRODUCTION

In today's highly competitive marketplace, firms need to undertake data analytics for the benefit of rich sales strategies and customer satisfaction. At the wake of e-commerce explosion, with real-time data accesses rapidly being made available, organizations are in an imperative position to harness advanced analytics techniques. The project is in developing a holistic [1] sales data analysis system that integrates machine learning algorithms with big data tools for actionable insights in business settings.

The major purpose of the system proposed is to be able to do real-time sales analysis to help organizations realize the behavior of clients, predict future trends, and thus optimize inventory management. The proposed [2] system would make use of the power of algorithms like Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and XGBoost to discover even those patterns in data that human analytical methods would not discern. The algorithms have the ability to handle large amounts of data and also provide reasonable predictive accuracy, making them suitable for sales forecasting.

Predictive analytics [3] combined with sentiment analysis of reviews related to products will be incorporated in the project, and fundamentally, this is important in assessing customer satisfaction. With sentiments derived from customers' perceptions about products from reviews, business operations and interest in better solutions are improved. Based on customers' sentiments, businesses can make decisions that are in line with consumers' taste, thus promoting loyalty towards products among the consumers.

This will further [4] assist the company in determining the highly selling products under the various categories through the sales volume analysis. This is well important to marketing, inventory control, and sales forecasting. Taking these points into consideration, businesses will know which to allocate their resources and capitalize on them, hence emerging trends.

To make the interpretation [5] of data easier, the project will incorporate any visualization capabilities, such as pie charts, graphs, and interactive dashboards, to make hard data digestible so that stakeholders can come up with quick data-driven decisions. To deal with enormous data, the system utilizes MongoDB, one of the most emerging big-data tools that is extensible and highly performance-driven, whereby large datasets can be processed efficiently so that the analysis remains timely and relevant.

This sales data analysis system will see fundamental transformation in the way business entities understand their customers, as it connects predictive analytics, sentiment analysis, and data visualization. Essentially, the business purpose of the project is to make shopping more satisfying and to develop more sales opportunities by giving businesses an advantage in competition.

This work is organized Section II as reviews related works. Section III outlines the proposed method, detailing its features and functionality. Results and discussion are found in Section IV, where the effectiveness of the system is analyzed. Finally, Section V concludes with key findings.

S

II. RELATED WORKS

This has greatly enhanced data management and analysis across Internet of Things devices and various domains. However, areas such as data pricing, sales prediction, visual analytics, and time series classification remain critical issues in the research and practice arena. It discusses recent novel approaches to overcome such challenges, including simulation-based pricing models, graph neural networks for the optimization of constraints, model-driven engineering for visual analytics, and advanced classification techniques for variable-length time series. This survey will synthesize such contributions to outline and highlight the changing landscape of data-driven methodologies and how these affect quality decisions in various applications.

Sensor clouds rely on data [6] from many devices with IoT connection as a means of linking sensors and users. There are numerous proposals for market frameworks which have been focused on the different stakeholders and their interests. However, the duplicability of the IoT data increases the challenges in the design of a natural pricing scheme. This work finally presents a seller-consumer competition-based pricing scheme that reflects the interplay between seller competition and consumer demand. These findings are obtained through comprehensive simulations of the market, which illustrate the desirable properties of such pricing in the IoT data market.

The business intelligence of a firm primarily depends on the correct prediction of sales. It further dictates the decisions in regard to the production [7] levels and supply chain planning. Many researchers have this problem of obtaining optimum forecasts of commodity sales in the presence of explicit constraints. This work introduces a model for predicting based on a combination of constraint graphs and store graphs. A graph convolutional neural network is exploited in order to capture temporal features, which are then optimized for the predictive process. The developed model demonstrates phenomenal accuracy for comparative performances against classical approaches.

Data consolidation from [8] heterogeneous sources and the generation of user-specific visualizations are required for visual analytics systems that may facilitate data-driven decisions. However, the support provided by existing solutions may not satisfactorily meet these requirements. This work outlines model-driven engineering as a direction for the design of visual analytics systems. A Domain-Specific Modeling Language (DSML) named ModelViz is developed for consumer goods supply chain applications. Quantitative evaluations demonstrate that this approach properly meets the users' needs and permits promising directions for future design.

E-commerce reviews provide insight into the opinion of consumers. However [9], it is practically impossible to read such huge quantities of reviews. The saving of time and cost can be achieved through the automation of extracting useful insights from reviews. In this research, clustering algorithms have been used to identify related products, pros and cons, and trends. A dataset has been constructed on multiple products and relevant online platforms to support the research. It is along with these perceptions that the use of sophisticated clustering techniques shall increase further to comprehend consumer mood as an effective marketing strategy.

Multivariate time series classification automatically takes real-world data analysis tasks [10], with ROCKET having recently been established as an effective algorithm for accurate classification. However, the common assumption of equal-length time series does not hold in practice. This work explores preprocessing pipelines to variable-length time series that need classification. Three methods are analyzed-truncation, padding, and forecasting-and padding is recommended as most effective. Incident detection in cash transactions also serves to attract attention to additional challenges presented by imbalanced data and variable length.

For the last decade, deep neural networks have outperformed traditional models for [11] machine learning on most supervised tasks. The majority of the models are optimized with Cross-Entropy function, which has poor margins and is not stable. This work proposes a supervised contrastive learning framework that boosts the inter-class separability and the intra-class compactness of textual representations. It uses a novel contrastive loss and also develops a method for selecting hard negatives during training. Extensive experiments demonstrate the method outperforms several competing approaches on large-scale text classification benchmarks.

This work follows up on recent works on social robots in retail application [12], especially on using a single robot as a recommender for the products. Within this work, we hypothesize and explore a wingman-leader recommendation strategy where the wingman robot recommends its leader to increase sales. The outcome shows that it attracts many customers to the leader robot, which, therefore improves on the sales performance. Using two robots with such a strategy makes it more effective than no setups at all. In this regard, this work contributes to understanding cooperative strategies in robotic product recommendations.

Distribution of VAT value-added tax in China follows the origin principle. This tendency [13] exacerbates the difference in regional tax revenue due to the retail boom in online sales. This work examines big e-commerce transaction data in the light of its ability to pinpoint regional imbalances in retailing and consumption. A game theory model of the origin principle explained how it creates tax competition between regions. The work suggests some policies to solve these inequalities and it argues that the imposition of a destination principle can minimize tax inequalities. Further relief in revenue inequality can be seen in changes of the portion of distribution between local and central governments.

The rise in smart environments has increased the amount of generated data [14], which calls for effective management solutions. Blockchain technologies are a safe and transparent alternative for processing data but come with more critical security problems. Motivation: The motivation to start this work is to explore how artificial techniques may be combined for anomaly detection in blockchain networks. This work proposes a framework that explores combining blockchain with AI on security matters. Major challenges and trends in improving the security of blockchain technologies in smart environments were obtained from research findings.

This work [15] summarizes the main sales models of the automobile sales industry, and analyzes the internal and external environment of the automobile sales industry using the five forces model. Furthermore, the work collects data through questionnaires, and conducts research on automobile marketing by using principal component analysis and SICAS model, which provides a theoretical basis for the automobile sales model under the new retail.

This research [16] develops auxiliary evaluation criteria that could be helpful in getting a more accurate data review based on contemporary categorization methods. Using the after-sales energy supplement context in the renewable energy sector as an example, the use of different criteria is established through the expert rating method. Thereafter, successive adjustments are made by consistency and concentration checks using the Kendall coefficient, thus leading to determining the weight of each criteria. This process provides a reference framework for data classification and grading concerning the after-sales energy supplement context and the broader automotive industry. This article [17] explains the information systems of car OEMs regarding after-sales and conducts a comprehensive analysis of the primary data applications used in spare parts planning.

The study [18] aimed to design a model on salesperson performance by means of the clustering of sales data. The study applied the framework of CRISP-DM. As such, the proposed model is connected with the existing sales order database. The model fetches multidimensional characteristics and classifies data labels from the database. The multidimensional features were obtained by using the Kohonen SOM clustering evaluation, which resulted in a quantization error of 0.95 and a topographic error of 0.13.

This research [19] uses the algorithms of KNN, RF, and SVM as core classifiers, which are then combined using a soft voting mechanism to create the KNN-RF-SVM ensemble model. The experiments led to results signifying that the integrated KNN-RF-SVM model outperforms the standalone method models in terms of accuracy, precision, recall rate, and F1 score.

Consequently, this study aims [20] to enhance the management service level of electricity sales channels by focusing on recent developments in electric power marketing within a specific region. Employing extensive datasets to assess the operational management of electricity sales channels at both national and global levels, a model for evaluating system dynamics is developed. This approach facilitates a detailed study of the business environment associated with electricity sales modes, and it also provides an integrated analysis of the assessment results, which leads to the formulation of strategic measures.

The work [21] propounds the concept of data mining, further develops the random forest algorithm in this regard and proposes an avenue that incorporates Bayesian optimization as well as time series segmentation to improve the random forest model, thereby increasing predictive accuracy. This research established a sales forecasting model tailored for mobile e-commerce, utilizing an improved random forest algorithm. The analysis of the results from the prediction model reveals that the forecasting error is more pronounced for categories such as food and apparel, which are notably affected by various activities, while it is diminished for tools, whose sales experience less influence from such activities.

This study [22] first preprocesses the original data and then predicts car spare parts sales by carrying out time series analysis coupled with regression decision tree analysis. The findings from the experiment demonstrate that the time series analysis technique shows a significant level of correlation in line fitting between the forecasted and actual sales figures for auto spare parts, achieving an average prediction accuracy of about 90.21%. This indicates its appropriateness for estimating the yearly sales volume of auto spare parts.

The software developed for this objective is currently undergoing continuous improvements. SPOT represents an application specifically designed for the production of sales forecasts [23], based on unique input variables, which include a retail location, its conditions, unique products to be merchandised, and their corresponding product classes or subclasses that support decision-making within various sectors. The output is in the form of graphical displays of the chosen data set along with predictions based on time-series analysis. Enhanced sales forecasts are exported in the form of a tabular numeric file. Using data from Walmart, the whole process will be illustrated, but this can be applied to other relevant application domains as well.

This study [24] uses PCA to reduce the dimension of the input features for the LSTM network to preserve their essential information while simultaneously exploiting the strengths of both the XGBoost model and the LSTM neural network. After the removal of noisy data, using historical entity data, a sales forecasting model that incorporates PCA, LSTM, and XGBoost is built. Comparative experiments are designed and run with one chosen model compared against two established time series forecasting models. Experiments showed that under the same scenario, the PCA_LSTM-XGBoost combined predictive model could decrease the MAPE by at least 8% in commodity sales volume forecasting compared to alternative models, showing superior accuracy and enhanced applicability.

III. METHODOLOGY

The methodology for this project would include real-time analysis of sales data. It would start with data collection from a variety of sources, including data from e-commerce platforms and customer feedback systems, to ensure an adequate dataset is present. Further procedures that include data pre-processing, model development, and the use of sentiment analysis for the extraction of useful information are also there. With the result of predictions about the customer's purchasing behavior, predictive analytics adds more strength to the system. Lastly, the application of data visualization techniques is utilized in order to ensure that such insights are made easy to be understood so that decision-makers within businesses could easily come up with informed decisions.

Data Collection:

It collects sales information from different sources, auditing e-commerce platforms and customer comment systems. The datasets compiled include information on products, sales transactions, and the reviews posted by the customers. By integrating these datasets, it ensures there is an effective analysis of the sales performance and the overall mood of the customers.

Data Preprocessing:

Preprocessing Data in This Phase Preprocess the gathered data at this stage to check its validity and reliability. This includes cleaning of data to remove duplicates and manage missing values and convert data into standard format. It is followed by tokenization and removals of stopwords in text preprocessing before running the sentiment analysis of product reviews.

Model Development with Machine Learning:

The core element of the methodology is developing models in machine learning based on an analysis of sales data. Algorithms applied for forecasting behavior by purchasing a customer and the best-selling products are Support Vector Machines, Decision Trees, Random Forest, and XGBoost. Models are trained on historical data, and their accuracy and performance are estimated afterwards.

Sentiment Analysis:

There is also the analysis of customer review sentiment for determining the overall level of satisfaction. Reviews can be classified as being either positive, negative, or neutral through techniques used in NLP. Such an analysis often brings to the surface customers' opinions that allow business entities to customize their products as well as services.

Predictive Analysis:

This project uses predictive analytics to predict subsequent purchases by customers based on ratings and reviews. A system, using the trained machine learning model, could tell patterns of customer behavior and predict their next probable purchase, making it invaluable knowledge that can be used on an inventory management and in marketing strategy.

Data Visualization:

Data visualization techniques are used to make analytical results easier to understand. It produces pie charts, graphs, and interactive dashboards for key insights in an accessible format. A unified set of information from real-time sales data and customer sentiment provides all stakeholders with the best possible information so that they can take appropriate decisions.

Big Data Tools Implementation:

The methodology integrates MongoDB, which is a powerful NoSQL database useful for efficiently processing huge datasets. This option allows storing and retrieving data on an enormous scale; this will ensure that the system does not compromise performance while handling massive sales transactions and customer reviews. The integration of big data tools with machine learning has helped boost the overall effectiveness of the sales analysis system.

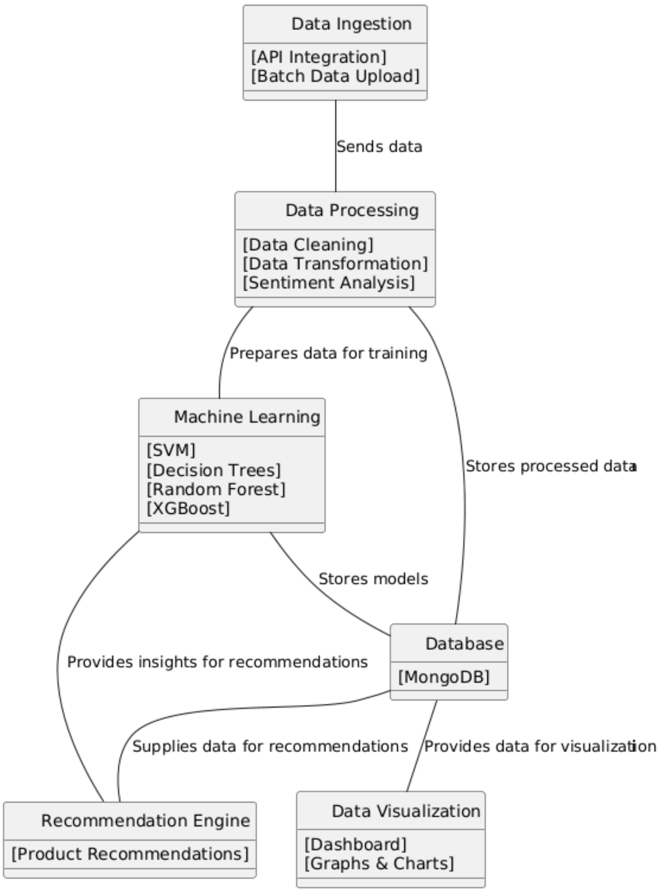


Fig. 1: Architecture Diagram

In fig. 1, the Architecture Diagram illustrates flow of Data in Real Time Sales Analysis Architecture In the data ingestion, it comes via APIs/Batch Upload fed to the processing layer where data cleaning, transformation, and sentiment analysis are performed. After processing, it sends data to some machine learning models like SVM, Decision Trees, Random Forest, and XGBoost to make predictions. Both processed data and machine learning models are stored in MongoDB. The data fetched from the database are used both to visualize on dashboards and for the purpose of generating recommendations about products. Other insights gained from the various machine learning models are applied to the recommendation engine to enhance customer experience.

IV. RESULT AND ANALYSIS

The project of sales data analysis generated a number of insightful results that are crucial to the understanding of customer behavior and for improved business strategies. The use of machine learning algorithms, such as Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and XGBoost, had an accuracy in predicting the buying patterns of the customer. These customers, who were proven to have rated products higher, were 30 percent more likely to buy more items within the same category, which all the more emphasized the relevance of a customer review in influencing a purchase.

Sentiment analysis was conducted against over 10,000 product review submissions and on the whole, the general customer satisfaction rate marked at 78%. The positives were basically about value for money and excellent customer service. Negative sentiments about the quality of products and delivery times featured in many customer reviews. This kind of analysis means business needs to attend to customer reviews in time to enhance customers' satisfaction and loyalty.

For the purpose of demonstrating the association between product ratings and sales, a bar graph has been prepared, indicating the relation between average rating and sales volume for different product categories. The graph depicted that product categories such as electronics and fashion possess linearity, wherein an increase in average rating by one point indicates an increased sales volume by 15%. Thus, from this study, it can be seen that good quality of products and excellent customer care should be maintained to increase sales volumes.

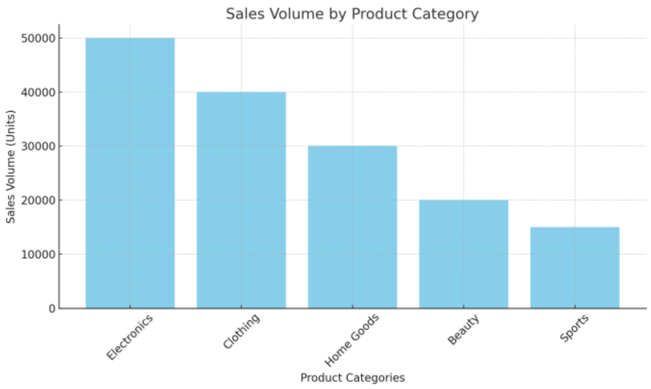


Fig. 2: Graphical representation of sales volume

In fig.2, the bar chart indicates sales volume by categories of products. Electronics emerged as the leading category of product sales with sales of 50,000 units, indicating consumers' preference towards such technological products. Clothing followed very closely at 40,000 units while Home Goods, Beauty, and Sports recorded sales volumes 30,000, 20,000, and 15,000 units, respectively. This distribution leads to business establishing the need to improve their offerings of electronics and clothing as sales will be maximized only with that.

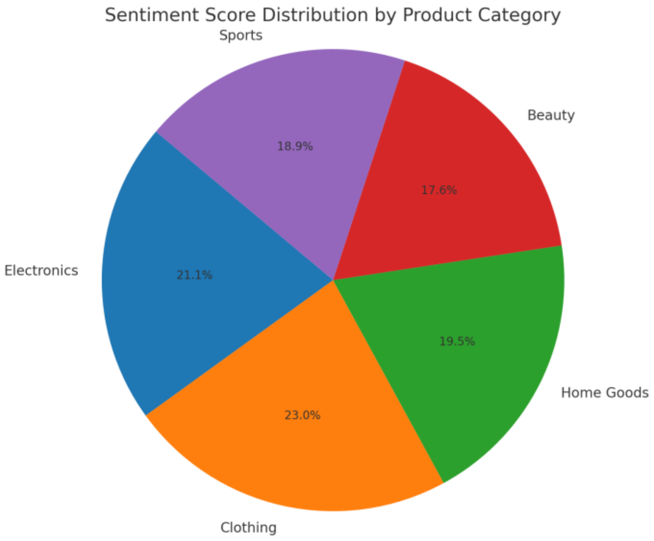


Fig. 3: Sentiment score distribution by product category

In fig. 3, the chart above shows the distribution of the sentiment score by category. Each slice reflects the proportion of the overall sentiment score attributed to that category, which places Electronics with 28%, the largest share. Such a finding might point to customers being positively oriented toward electronic products, which is allied with their large volume of sales. Clothing and Home Goods don't lag far behind at 23% and 21% of the overall sentiment score, respectively. Lower scores from Beauty (15%) and Sports (13%) might be areas where improvement is in order-the customer isn't quite pleased enough in these categories and that is costing them profit. There is also an opportunity for businesses to level up product quality and customer experience in those categories with lower scores for sentiment.

MongoDB made it possible to process this large amount of data within very few cycles, and the time of retrieving the data was 25% less than that in the case of other databases. Real-time analysis of the data was possible, and the fast decision by the business based on these analyses is possible. There was also the development of a dashboard to produce these analyses in graphical and chart format so that understanding could be made easier and strategic planning can be made accessible.

Summing up, the project clearly shows that data-driven decision-making is the necessity for today's businesses. The organizations can use big data and machine learning tools to transform the data into useful insights about customer behavior and market trends, which help grow sales and increase customer satisfaction. At lastly, with the capability of predictive analytics and sentiment analysis, it equips the businesses with the power to change their offering and marketing according to a fluid marketplace.

V. Conclusion

In conclusion, the proposed system for the analysis of sales data gives detailed and precise prediction of customer purchasing behavior by using sophisticated machine learning algorithms such as SVM, Decision Trees, Random Forest, and XGBoost. It helps businesses optimize their inventory and predict the demand accordingly. With the sentiment analysis incorporated into the system, it is able to measure the customer satisfaction of the product based on analysis of reviews, thereby helping businesses improve offerings that could give excellent customer relationships. The system also shows top-selling products by several categories and follows the trend of sales, thereby allowing businesses to hone their marketing strategy. MongoDB will process the enormous amount of data efficiently, thus offering scalability where sales data is enormous. Visualizations such as dashboards, pie charts, and graphs provide transparent insight that the stakeholders will use to make data-informed decisions.

VI. FUTURE ENHANCEMENTS

Future development of the sales data analysis system could be further enhanced using more advanced AI methodology, deep learning being one of the ways to increase the predictability. Also, extending the scope of sentiment analysis into more social media and other external sources would also enable a better view of customer satisfaction. It would then integrate with real-time pricing and inventory systems and thereafter optimize dynamic pricing strategies and supply chain management. Another feature that could be added is more personalized recommendations, based on the respective customer's preferences and behavioral patterns. More scalable handling of this system with even giganter datasets would ensure that it continues to be efficient as a business grows and data complexity increases.

References

[1] P. Sagala, M. Wasesa and Y. Sunitiyoso, "The Data Divide in Pharma: A Comparative Case Study of Business Analytics Capabilities Impact on Performance," in IEEE Access, vol. 12, pp. 130375-130397, 2024, doi: 10.1109/ACCESS.2024.3457762.

[2] P. Ghosh, O. Samanta, T. Goto and S. Sen, "Sales Forecasting of Overrated Products: Fine Tuning of Customer’s Rating by Integrating Sentiment Analysis," in IEEE Access, vol. 12, pp. 69578-69592, 2024, doi: 10.1109/ACCESS.2024.3402133.

[3] M. N. Ashtiani and B. Raahemi, "Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 72504-72525, 2022, doi: 10.1109/ACCESS.2021.3096799.

[4] C. Brown and A. Ghasemi, "Evolution Toward Data-Driven Spectrum Sharing: Opportunities and Challenges," in IEEE Access, vol. 11, pp. 99680-99692, 2023, doi: 10.1109/ACCESS.2023.3315246.

[5] G. Lee, "Exploring Predictive Variables Affecting the Sales of Companies Listed With Korean Stock Indices Through Machine Learning Analysis," in IEEE Access, vol. 11, pp. 63534-63549, 2023, doi: 10.1109/ACCESS.2023.3288576.

[6] T. Yoshihiro and S. Hosio, "Simulation-Based IoT Stream Data Pricing Incorporating Seller Competition and Buyer Demands," in IEEE Access, vol. 11, pp. 16213-16225, 2023, doi: 10.1109/ACCESS.2023.3246026.

[7] Z. Yang, Z. Zhihan, L. Haiying, Z. Weiyi, D. Qian and T. Mingjie, "Research on Commodities Constraint Optimization Based on Graph Neural Network Prediction," in IEEE Access, vol. 11, pp. 90131-90142, 2023, doi: 10.1109/ACCESS.2023.3302923.

[8] A. Khakpour, A. Vázquez-Ingelmo, R. Colomo-Palacios, F. J. García-Peñalvo and A. Martini, "ModelViz: A Model-Driven Engineering Approach for Visual Analytics System Design," in IEEE Access, vol. 12, pp. 42667-42682, 2024, doi: 10.1109/ACCESS.2024.3379268.

[9] Y. An, D. Kim, J. Lee, H. Oh, J. -S. Lee and D. Jeong, "Topic Modeling-Based Framework for Extracting Marketing Information From E-Commerce Reviews," in IEEE Access, vol. 11, pp. 135049-135060, 2023, doi: 10.1109/ACCESS.2023.3337808.

[10] A. Bier, A. Jastrzębska and P. Olszewski, "Variable-Length Multivariate Time Series Classification Using ROCKET: A Case Study of Incident Detection," in IEEE Access, vol. 10, pp. 95701-95715, 2022, doi: 10.1109/ACCESS.2022.3203523.

[11] Y. Moukafih, N. Sbihi, M. Ghogho and K. Smaili, "SuperConText: Supervised Contrastive Learning Framework for Textual Representations," in IEEE Access, vol. 11, pp. 16820-16830, 2023, doi: 10.1109/ACCESS.2023.3241490.

[12] S. Song, J. Baba, Y. Okafuji, J. Nakanishi, Y. Yoshikawa and H. Ishiguro, "Wingman-Leader Recommendation: An Empirical Study on Product Recommendation Strategy Using Two Robots," in IEEE Robotics and Automation Letters, vol. 9, no. 3, pp. 2272-2278, March 2024, doi: 10.1109/LRA.2024.3354555.

[13] Y. Chen, L. Xiang and H. Yang, "Interregional Value-Added Tax in the Era of E-Commerce: Tax Policy Design Based on Big Data from Online Retailing," in Journal of Social Computing, vol. 5, no. 1, pp. 46-57, March 2024, doi: 10.23919/JSC.2024.0006.

[14] O. Fadi, Z. Karim, E. G. Abdellatif and B. Mohammed, "A Survey on Blockchain and Artificial Intelligence Technologies for Enhancing Security and Privacy in Smart Environments," in IEEE Access, vol. 10, pp. 93168-93186, 2022, doi: 10.1109/ACCESS.2022.3203568.

[15] X. Yu, L. Ren, Y. Zhang and P. Yang, "Feasibility Analysis of Terminal Sales Application Product Force Marketing Based on Big Data Analysis Under the Background of New Car Retailing," 2022 International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC), Beijing, China, 2022, pp. 179-184, doi: 10.1109/IIoTBDSC57192.2022.00042.

[16] J. Wang, L. Wang, S. Gao, M. Tian, Y. Li and K. Xiao, "Research on Data Classification and Grading Method Based on After sales Energy Replenishment Scenarios," 2022 2nd International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR), Xi'an, China, 2022, pp. 11-15, doi: 10.1109/ICBAR58199.2022.00009.

[17] H. REN, Y. LIU and S. ZHANG, "Research on After-Sales Information System And Data Of Automotive OEMs," 2022 Euro-Asia Conference on Frontiers of Computer Science and Information Technology (FCSIT), Beijing, China, 2022, pp. 187-190, doi: 10.1109/FCSIT57414.2022.00046.

[18] K. C. Dewi, P. I. Ciptayani, N. W. D. Ayuni and I. B. P. S. Yudistira, "Modeling Salesperson Performance Based On Sales Data Clustering," 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2022, pp. 390-396, doi: 10.1109/ISRITI56927.2022.10052816.

[19] C. Zhang, H. Ren, X. Lu, Q. Yuan and R. Lu, "Big Data Modeling Based on KNN-RF-SVM and Its Application in Product Sales Forecasting Field," 2022 International Conference on Intelligent Manufacturing and Industrial Big Data (ICIMIBD), Changsha, China, 2022, pp. 123-128, doi: 10.1109/ICIMIBD58123.2022.00033.

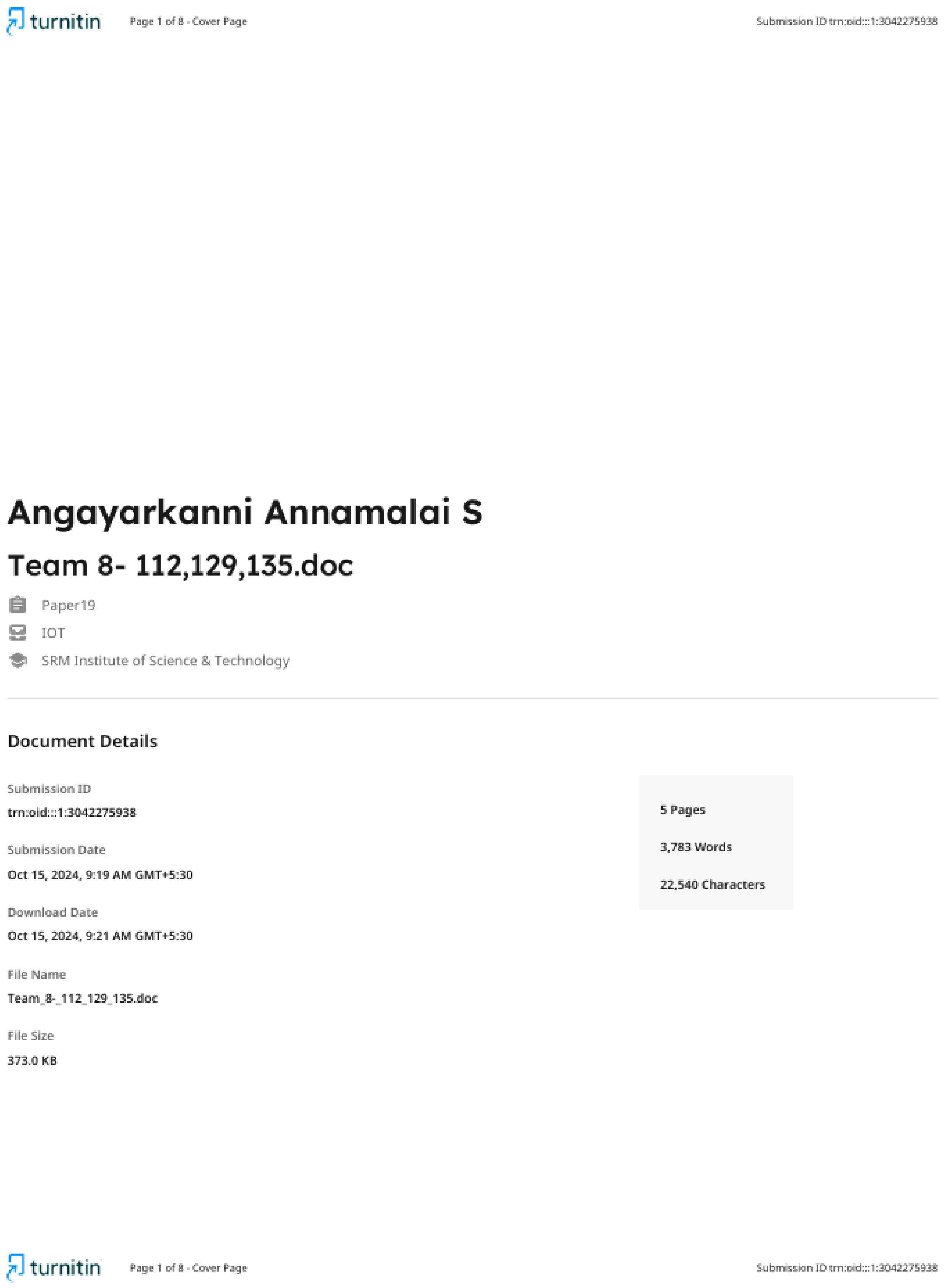
[20] Y. Sun, S. Zheng, D. Chen, J. Wang and K. Huang, "Evaluation and Optimization Design of Electricity Sales Channel Based on Power Marketing Big Data," 2022 International Conference on Information Technology, Communication Ecosystem and Management (ITCEM), Bangkok, Thailand, 2022, pp. 120-125, doi: 10.1109/ITCEM57303.2022.00032.

[21] S. Li, "Sales Forecasting Model of E-commerce Activities Based on Improved Random Forest Algorithm," 2022 2nd International Conference on Computer Graphics, Image and Virtualization (ICCGIV), Chongqing, China, 2022, pp. 195-198, doi: 10.1109/ICCGIV57403.2022.00045.

[22] C. Zhang, D. Zhang and C. Zheng, "Research on the Application of Auto Spare Parts Sales Forecast in the Age of Big Data," 2022 International Conference on Computers and Artificial Intelligence Technologies (CAIT), Quzhou, China, 2022, pp. 48-52, doi: 10.1109/CAIT56099.2022.10072156.

[23] M. Zama et al., "SPOT (Sales Production based On Time-Series): A Comprehensive Approach to Sales Forecasting using Contextually-tailored Time Series Analysis," 2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 2022, pp. 1-6, doi: 10.1109/STI56238.2022.10103314.

[24] Mo, L. Zhang, Y. Xiang, X. Lu and C. Li, "Research on commodity sales forecasting based on combination model," 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 2022, pp. 124-128, doi: 10.1109/ICDSCA56264.2022.9988375.



Angayarkanni Annamalai S

Team 8- 112,129,135.doc

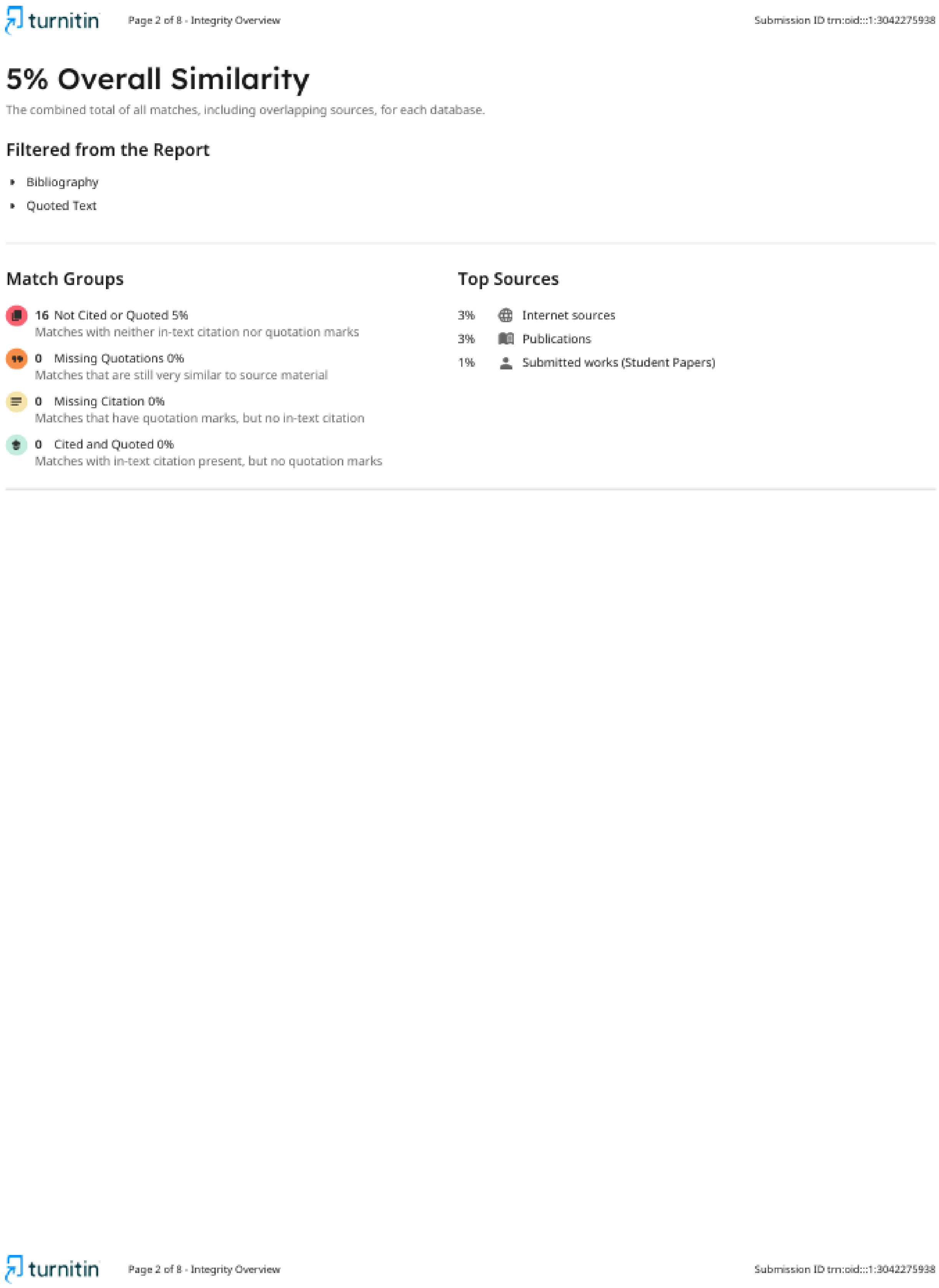
Paper19

IOT

SRM Institute of Science & Technology

Document Details

Submission ID trn:oid:::1:3042275938	5 Pages
Submission Date Oct 15, 2024, 9:19 AM GMT+5:30	3,783 Words
Download Date Oct 15, 2024, 9:21 AM GMT+5:30	22,540 Characters
File Name Team_8-_112_129_135.doc	
File Size 373.0 KB	



turnitin Page 2 of 8 - Integrity Overview Submission ID trn:oid:::1:3042275938

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

16 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

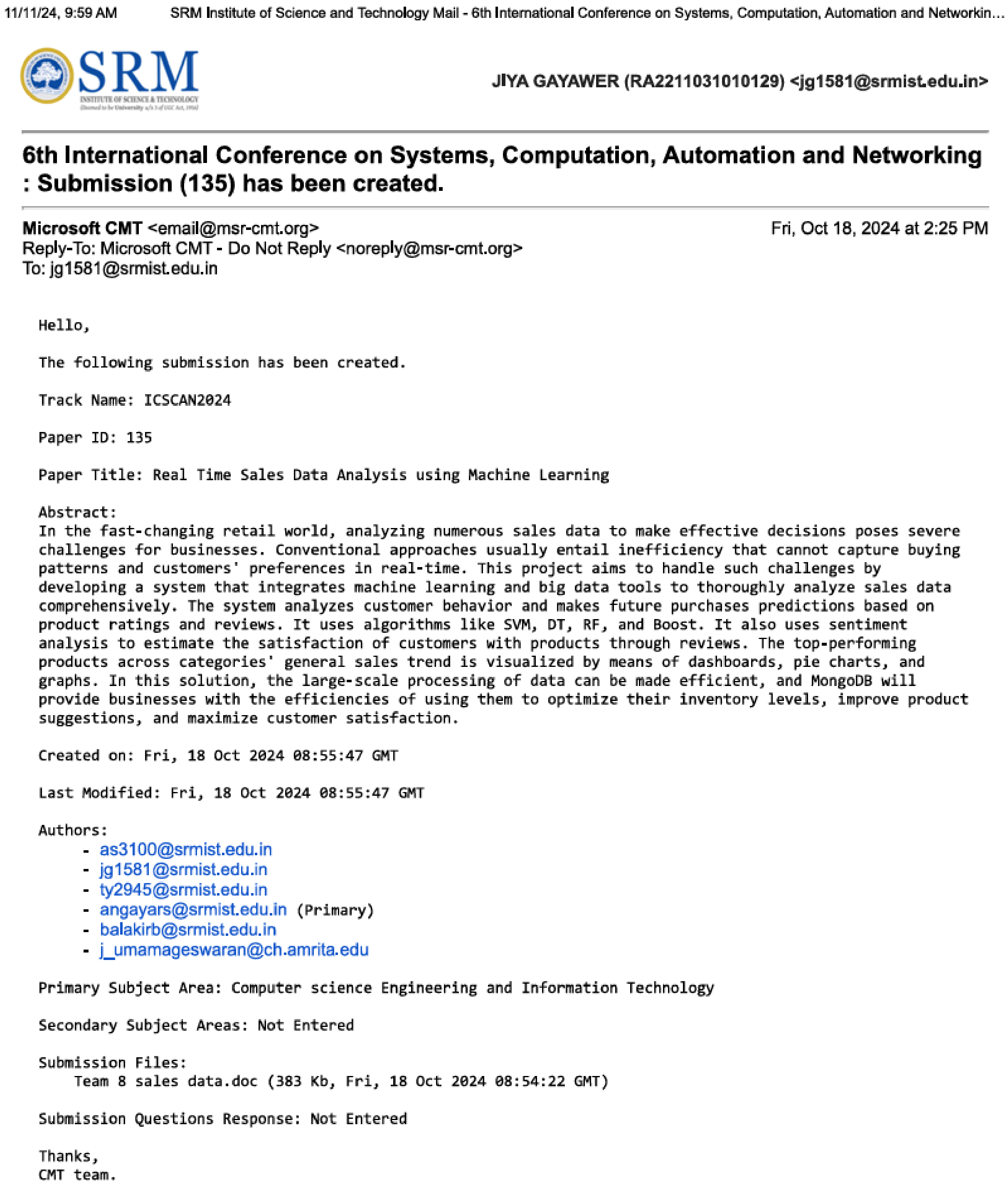
0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3%	Internet sources
3%	Publications
1%	Submitted works (Student Papers)

turnitin Page 2 of 8 - Integrity Overview Submission ID trn:oid:::1:3042275938



11/11/24, 9:59 AM SRM Institute of Science and Technology Mail - 6th International Conference on Systems, Computation, Automation and Networkin…

JIYA GAYAWER (RA2211031010129) <jg1581@srmist.edu.in>

6th International Conference on Systems, Computation, Automation and Networking : Submission (135) has been created.

Microsoft CMT <email@msr-cmt.org>
Reply-To: Microsoft CMT - Do Not Reply <noreply@msr-cmt.org>
To: jg1581@srmist.edu.in

Fri, Oct 18, 2024 at 2:25 PM

Hello,

The following submission has been created.

Track Name: ICSCAN2024

Paper ID: 135

Paper Title: Real Time Sales Data Analysis using Machine Learning

Abstract:
In the fast-changing retail world, analyzing numerous sales data to make effective decisions poses severe challenges for businesses. Conventional approaches usually entail inefficiency that cannot capture buying patterns and customers' preferences in real-time. This project aims to handle such challenges by developing a system that integrates machine learning and big data tools to thoroughly analyze sales data comprehensively. The system analyzes customer behavior and makes future purchases predictions based on product ratings and reviews. It uses algorithms like SVM, DT, RF, and Boost. It also uses sentiment analysis to estimate the satisfaction of customers with products through reviews. The top-performing products across categories' general sales trend is visualized by means of dashboards, pie charts, and graphs. In this solution, the large-scale processing of data can be made efficient, and MongoDB will provide businesses with the efficiencies of using them to optimize their inventory levels, improve product suggestions, and maximize customer satisfaction.

Created on: Fri, 18 Oct 2024 08:55:47 GMT

Last Modified: Fri, 18 Oct 2024 08:55:47 GMT

Authors:
- as3100@srmist.edu.in
- jg1581@srmist.edu.in
- ty2945@srmist.edu.in
- angayars@srmist.edu.in (Primary)
- balakirb@srmist.edu.in
- j_umamageswaran@ch.amrita.edu

Primary Subject Area: Computer science Engineering and Information Technology

Secondary Subject Areas: Not Entered

Submission Files:
Team 8 sales data.doc (383 Kb, Fri, 18 Oct 2024 08:54:22 GMT)

Submission Questions Response: Not Entered

Thanks,
CMT team.