

METHODOLOGY TO MINIMIZE RISK IN SME LENDING FOR THE US AND  
INDIA MARKETS

by

Neeraj Mehta, Bachelor of Technology

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

September, 2024

METHODOLOGY TO MINIMIZE RISK IN SME LENDING FOR THE US AND  
INDIA MARKETS

by

Neeraj Mehta

Supervised by

Dr. Kamal Malik

APPROVED BY

*Nihar Behera*

---

Dr. Nihar Behera - Dissertation Chair

RECEIVED/APPROVED BY:

---

Admissions Director

## **Dedication**

This is dedicated to all the lenders in a financial market who would like to lend SME population or the segment of the new businesses of the world to enable them to expand their businesses and contribute to the economic growth of the society in a broader sense. It is hoped that this work would be helpful in some way to the exploration in the field of small and medium business lending. I hope it will be useful to them in navigating the risk involved in financing to small and medium-sized businesses.

## **Acknowledgements**

I would want to use this opportunity to convey my deep thanks and debt of gratitude to Dr. Nirmala Srivastava, my internal guide, for giving me all the support and direction I required. Their generous fellowship has enabled me to undertake this doctorate program at Swiss School of Business and Management Geneva. Their insightful recommendations were really beneficial and gave me a wealth of knowledge. The Subject Matter Experts in Retail and SME Lending provided help and direction for this study. I am very grateful to my mentor, Dr. Kamal Malik. Their feedback, knowledge, and experience have significantly aided the study and enhanced the report. My gratitude also goes out to the Swiss School of Business and Management Geneva for giving me the chance to conduct this research, which is extremely important in the fast-paced environment for economic growth and the survival of new firms. Thankful to Dr. Devashree Jadhav for her insightful comments on the study. Above all, I want to thank my mother and wife from the bottom of my heart. Their encouragement gave me the bravery to overcome numerous obstacles that I have encountered during my study.

ABSTRACT

METHODOLOGY TO MINIMIZE RISK IN SME LENDING FOR THE US MARKET

Neeraj Mehta  
2024

Dissertation Chair: Dr. Nihar Behera  
Co-Chair: <If applicable. Co-Chair's Name>

Fintech start-ups offer digital SME Lending Platforms for smooth customer journeys, that spans from campaigning to loan delivery to small enterprises. These startups are rising in numbers across the globe specifically the United States and India Markets due to the increased demand of the credit by the small and medium enterprises. The widely used phrase "lending" refers to the process of lender giving money to the borrower. Digitalization uses state through processing to bring cost-effectiveness, enhanced customer experience, speed of disbursements, better risk assessment, and adequate decision-making. SME, or small and medium-sized enterprises, are companies that need capital to operate in the market. To grow their business, they require capital. Since the majority of them are new companies, they fall into the subprime or thin-file consumer category, which means that credit agencies have either given them a low credit score or no credit score at all. These customers' loan requests are denied by financial institutions because they are unable to meet their capital criteria. By leveraging alternative data such as payments, rent, mobile phone, and social media information, the fintech start-up industries target these customers and offer loans at higher interest rates in an effort to reduce the risk. They have also developed their own digital systems for risk assessment of these customers.

This paper provides the improvisation pertaining to the Bank Statement Analyser built in the study (Mehta, 2020) to mitigate the risk associated with the SME lending in the United States and India. At the same time this method uses the latest AI technology via Large Language Models that has a potential to boost the credit decisioning process of a financial institution.

The statements that banks periodically release are known as bank statements. Transaction-level information for their clients is included in these statements. This study examined 1,00,000 transactions from 3,300 small and medium-sized businesses in the US and 20,000 transactions from 475 small and medium enterprises in the India market using anonymized parsed PDF bank statement data. Using text mining techniques like count vectorizer and n-grams, as detailed in the study (Mehta, 2020), this transaction-level data was first used to

generate transaction categories, or keywords. Later, these keywords were used to classify the transactions. Compared to earlier research (Mehta, 2020), a more refined set of keywords is used in this study. In this work, rule-based, machine learning, and Large Language Models methods for multi-class classification were employed as classification techniques. Various prompt designs and strategies are used for large language models classification task.

In the end, a number of important measures known as cash flow variables were generated for loan credit decisioning, together with changes in revenue, operational profit, operating profit margin, and debt services.

**KEYWORDS:** LLM in Banking, Bank Statements Analysis, Machine Learning in Lending, Digital SME Lending, Digital Lending, Risk in SME Lending, AI to reduce risk in SME Lending, Use of large language models for credit decisioning.

## TABLE OF CONTENTS

List of Tables .....	8
List of Figures .....	9
LIST OF ABBREVIATIONS.....	10
CHAPTER I: INTRODUCTION.....	2
1.1 Introduction.....	2
1.1.1 Understanding SME Lending.....	3
1.1.2 Risk in SME Lending.....	5
1.1.3 Challenges associated with SME Lending.....	9
1.2 Research Problem .....	10
1.3 Purpose of Research.....	11
1.4 Significance of the Study .....	11
1.5 Research Purpose and Questions .....	16
CHAPTER II: REVIEW OF LITERATURE .....	17
2.1 Introduction.....	17
2.2 Digitalization in SME Lending.....	17
2.3 Understanding Cash Flow analysis.....	25
2.4 Methods used in past for Text Classification or Bank transactions classification .....	30
2.5 Problem statement.....	42
2.6 Gaps identified in the Literature Review .....	46
2.7 Summary .....	56
CHAPTER III: METHODOLOGY .....	58
3.1 Overview of the Research Problem .....	58
3.2 Operationalization of Theoretical Constructs .....	61
3.3 Research Purpose and Questions .....	61
3.4 Research Design.....	62
3.4.1 Elaboration to Approaches.....	62
3.4.2 Logical Flow Diagrams.....	64
3.4.3 Tools and Technologies .....	68
3.5 Population and Sample .....	70
3.6 Participant Selection .....	70
3.6.1 Data Preparation.....	70
3.6.2 Data Preparation for Keywords Generation Module.....	71
3.6.3 Data Preparation for Clustering and Rule-Based Classification Module .....	71
3.6.4 Data Preparation for Machine Learning Classification.....	71



3.6.5	Data Preparation for Large Language Model Classification.....	72
3.6.6	Data Preparation for Post Classification Module.....	76
3.7	Data Partitioning.....	76
3.8	Instrumentation.....	77
3.8.1	Transactions Clustering.....	77
3.8.2	Transactions Classification using Machine Learning.....	79
3.8.3	Transactions Classification using Large Language Model.....	80
3.8.4	Evaluation.....	80
3.8.5	Model Implementation.....	82
3.9	Data Collection Procedures.....	82
3.10	Data Analysis.....	84
3.10.1	Data Pre-processing.....	84
3.10.2	Data Transformation.....	85
3.10.3	Data Cleaning.....	87
3.10.4	Fixing Data Types.....	88
3.10.5	Exploratory Data Analysis.....	88
3.11	Research Design Limitations.....	99
3.12	Conclusion.....	100
CHAPTER IV: RESULTS.....		102
4.1	Research Question One.....	102
4.1.1	Keywords Generation Process.....	102
4.1.2	Keywords validations.....	111
4.2	Research Question Two.....	111
4.2.1	Machine Learning Classification.....	112
4.2.2	Machine Learning based classification Evaluation.....	112
4.3	Research Question Three.....	112
4.3.1	LLM Classification Results.....	113
4.3.2	Post Classification Results.....	124
4.3.3	Output Variables.....	124
4.4	Research Question Four.....	125
4.5	Summary of Findings.....	128
CHAPTER V: DISCUSSION.....		129
5.1	Discussion of Results.....	129
5.2	Discussion of Research Question One.....	129
5.3	Discussion of Research Question Two.....	130
5.4	Discussion of Research Question Three.....	130
5.5	Discussion of Research Question Four.....	131
CHAPTER VI: SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS.....		134
6.1	Summary.....	134

6.2	Implications.....	136
6.3	Recommendations for Future Research.....	137
6.4	Conclusion .....	139
	REFERENCES .....	142
	APPENDIX A: ENHANCED KEYWORDS LIST.....	150

LIST OF TABLES

Table 1 Gaps identified in Literature Review..... 46

## LIST OF FIGURES

Figure 1 Process Flow Diagram.....	65
Figure 2 Machine Learning Model/LLM Model .....	67
Figure 8 DBSCAN clustering process. (DBSCAN - Wikipedia, 2020) .....	78
Figure 9 Term Frequency as defined in (Maklin, 2019).....	86
Figure 10 Inverse Document Frequency as described in (Maklin, 2019).....	87
Figure 15 Number of Bank Accounts of Merchants.....	89
Figure 16 Number of Months of each Bank Account.....	90
Figure 17 Analysis of Credits, Debits & Monthly Averages of Bank Accounts .....	91
Figure 18 Analysis of Credits, Debits & Monthly Averages of Bank Accounts Part 2 ...	92
Figure 19 Merchant Level Analysis (1/2).....	93
Figure 20 Merchant Level Analysis (2/2).....	94
Figure 21 Industry Level Analysis.....	94
Figure 22 Bank ID Level Analysis .....	95
Figure 23 Visualization of Relationship .....	96
Figure 24 Top 5 Deposits of Merchants .....	96
Figure 25 Top 5 Deposits of Merchants part 2 .....	97
Figure 26 Number of Days between Withdrawals.....	97
Figure 27 Covariance Analysis.....	98
Figure 28 Cluster of similar transactions .....	99
Figure 29 Sample Keyword List.....	111
Figure 30 Set of output variables – LLM Model .....	125

## LIST OF ABBREVIATIONS

SME.....	Small and Medium Enterprises
US.....	United States
BSA.....	Bank Statement Analyser
AI.....	Artificial Intelligence
DSCR.....	Debt Service Coverage Ratio
DTI.....	Debt to Income Ratio
CFR.....	Cash Flow Ratio
CB.....	Credit Bureau
EDA.....	Exploratory Data Analysis
NBFC.....	Non-Banking Financial Companies
NASSCOM.....	The National Association of Software and Services Companies
UK.....	United Kingdom
PDF.....	Portable Document Format
NSF.....	Non-Sufficient Funds
EBITDA.....	Earnings Before Interest, Taxes, Depreciation, and Amortization
P&L.....	Profit and Loss statements
LLM.....	Large Language Model

CHAPTER I:  
INTRODUCTION

**1.1 Introduction**

Rao et al. (2023) had stated in their study “Small and medium-sized enterprises (SMEs) are a significant source of employment for many people in developing and developed economies. However, one of the significant constraints that SMEs face is limited access to finance.” Current study is all around the methods that could be adopted to reduce risk in SME lending and thus enable finance institutions to provide loans to the SME population for the development of their businesses that could further boost an economic growth of a country.

Over the past ten years, the digitalization of the lending industry has given rise to numerous fintech start-ups worldwide. These financial institutions are now engaged in tough competition. In the past, underwriters have made loan eligibility decisions for clients mostly based on credit agency credit scores. Because of this, a large percentage of customers have their credit demand denied by banks or are not eligible for loans because of low or no credit scores. These customers are referred to as subprime and thin-file customers. Fintech startups are focusing on this group of customers that either had their loan applications denied by banks or they required loans that could be processed more quickly for pressing demands.

These businesses take realistic risks into account and meet customer needs by offering unsecured (collateral-free) loans. Most of these loans have short terms and exorbitant interest rates. These fintech companies raise funds from investors and use it to make investments by lending money to customers. Commission fees from investors and various expenses related to approving, processing, and overseeing loans are their sources of revenue.

Appropriate risk measurement techniques and the use of alternative data sources are always needed in order to produce important risk metrics that could aid in auto decisioning for improved decision management. Only SME enterprises in the US and India markets were taken into account for this study, as they represent a significant portion of the customer base and are in continual need of funding to meet their working capital needs. The majority of them lack a credit score and credit history. In this study, bank statements that SMEs submit are examined to determine whether or not they qualify for a loan.

This study explores traditional and new AI technologies to boost the credit decisioning process. Rule based engine, Machine Learning and Large Language Models are used for credit decisioning of the SME consumers. Extensive Prompt Engineering is used in this study for the prompt designing in order to perform multi class classification using Large Language Model. In addition to this, few shot examples technique is employed for the LLM classification.

### **1.1.1 Understanding SME Lending**

Small and medium-sized enterprises (SMEs), which are essential to innovation, employment creation, and economic prosperity, are the cornerstone of economies. For many SMEs, obtaining funding is still a constant obstacle, especially in emerging nations and places of extreme poverty. Fintech companies must comprehend the financing environment for small and medium-sized enterprises (SMEs) because it has a direct impact on the viability and sustainability of these companies. SME financing includes a wide range of financial services and products that are designed to satisfy the capital requirements of small and medium-sized enterprises, from conventional bank loans to cutting-edge fintech options. On a national and worldwide level, these finance choices are essential for encouraging entrepreneurship, business growth, and economic development.

The SME lending landscape is defined by a broad range of opportunities and limitations, including market dynamics, regulatory frameworks, technological advancements, and

socioeconomic conditions. For an extended period, small businesses have mostly gotten their credit from commercial banks and microfinance institutions, two reputable sources of funding for small businesses. However, challenges like stringent collateral requirements, prolonged approval processes, and minimal outreach to minority SME market segments are often faced by these banks. Furthermore, the global financial crisis of 2008 worsened credit constraints for SMEs since banks tightened lending standards and reduced risk exposure in response to heightened regulatory scrutiny and economic uncertainty.

The financial technology industry, or fintech, emerged in recent years and transformed the SME lending sector by offering innovative solutions to numerous issues associated with traditional lending channels. Fintech companies leverage cutting edge technologies such as blockchain, AI, and data analytics to enhance credit evaluation capabilities, accelerate loan approval processes, and expand small and medium-sized enterprises' access to capital. By using automated algorithms and other data sources, fintech lenders can target underserved SME market segments that were previously kept out of conventional finance channels, make lending decisions more quickly, and offer more customized loan solutions.

The use of digital platforms and online lending has simplified SME financing, making cash acquisition for business owners relatively simple. Instead of the weeks or months that traditional lenders usually require, these platforms allow SMEs to complete loan applications online, submit documentation electronically, and receive funding within days or even hours. Furthermore, fintech lenders can interact with a wider range of SME clients—including startups, microenterprises, and companies with little collateral or credit history—by using risk-based pricing models and flexible underwriting guidelines.

In addition to making finance accessible, fintech-driven SME financing offers several significant advantages, including greater client satisfaction, lower transaction costs, and more transparency. Fintech platforms detect fraudulent behavior, more accurately assess credit risk, and customize loan terms to the borrower's needs and financial profile using



machine learning algorithms and data analytics. This data-driven approach's enhancement of creditworthiness assessment and decrease in default and non-performing loan incidence yield long-term benefits for lenders and borrowers alike. Fintech has the potential to revolutionize SME finance, but before this promise can be fulfilled, a variety of risks and restrictions must be taken into account. Fintech lenders find it very challenging to comply with regulatory standards, especially in highly regulated nations with strict licensing, consumer protection, and anti-money laundering (AML) regulations. Concerns around cybersecurity, algorithmic bias, and data privacy also highlight the significance of moral behavior and strong risk management frameworks for fintech-driven lending. Building trust and confidence in fintech lending platforms and supporting sustainable growth in the SME finance ecosystem requires adhering to responsible lending practices, treating borrowers fairly, and fully disclosing terms and conditions.

Globally, SME lending is essential for encouraging business, financial inclusion, and faster economic growth. The coming together of market dynamics, regulatory reforms, and technological innovation is creating new avenues for fintech-driven lending solutions to meet small businesses' financing requirements. Stakeholders in the SME ecosystem can fully leverage fintech's potential to support business people, stimulate innovation, and create robust economies that succeed in the digital age through data-driven insights, digital platforms, and collaborative partnerships.

### **1.1.2 Risk in SME Lending**

Lending to the SME population has advantages as well as disadvantages. Lending to SMEs carries a number of risks. When lending to SMEs, credit risk is a big concern. The financial standing of the SME, their capacity to repay loans, industry-specific risks, and demographic variables are some of the elements that contribute to credit risk. The term interest rate risk describes the swings in interest rates that might have an effect on borrowers who are SMEs. In order to maintain stability in their loan portfolios, fintech companies must successfully manage this risk. Variations in income might cause cash flow

issues, which can hinder their capacity to make loan payments on schedule. It is possible to define a suitable loan line for SMEs. Fintech must find a balance between controlling risk and providing the funding that SMEs require. SMEs frequently don't have much collateral.

Small and Medium-sized Enterprises (SMEs) are an essential component of the global economy, making a substantial contribution to economic growth, innovation, and the creation of jobs. However, due to a number of reasons, including lower financial transparency, fluctuating cash flows, and higher default probabilities compared to bigger firms, lending to SMEs has inherent risks for financial institutions and investors. Lenders, legislators, and business owners must all be aware of the risks involved in SME lending in order to implement risk-reduction plans, responsible credit evaluation procedures, and well-informed decision-making that promote ethical lending and long-term company expansion.

The intrinsic features of small enterprises, such as their limited financial resources, operational vulnerabilities, and reliance on important persons or market niches, pose a significant risk when it comes to financing to SMEs. SMEs frequently have difficulties that can affect their capacity to repay loans and fulfil financial commitments, such as limited access to formal finance, unstable market conditions, and regulatory concerns. Lenders may view SME financing as intrinsically riskier due to the absence of collateral or credit history, which makes it riskier than lending to large, established firms with track records and real assets.

Credit risk, which includes the possibility of borrower default or failure to repay loans in accordance with contractual commitments, is a major worry in SME lending. Factors like cash flow mismatches, market downturns, competitive challenges, or unanticipated events like natural catastrophes or economic crises can all cause financial difficulties for SMEs. Furthermore, lenders may be exposed to increased credit risk in the absence of strong frameworks for credit risk assessment and risk management, which might have a negative

effect on capital adequacy and profitability as well as increase the rate at which loans become delinquent and write off.

Operational risk, which arises from internal procedures, systems, and human mistake and may cause monetary losses or harm to financial institutions' reputations, is another important aspect of risk in SME lending. Inadequate underwriting criteria, poor loan paperwork, inefficient loan monitoring procedures, or insufficient controls to stop fraud and misconduct are a few examples of operational hazards in SME lending. Furthermore, operational hazards can be made worse by depending too much on antiquated technology, dispersed data systems, and manual processes, which can impair SME lending operations' scalability, efficiency, and regulatory compliance.

For SME lenders, market risk presents extra difficulties, especially in erratic or unpredictable economic climates where the creditworthiness and financial performance of SME borrowers can be impacted by macroeconomic variables, industry dynamics, and market conditions. The ability of SMEs to service debt and repay loans can be impacted by fluctuations in interest rates, exchange rates, commodity prices, and consumer demand. These fluctuations can also have an impact on SMEs' revenues, expenses, and profitability. Furthermore, new market uncertainties and systemic risks may be introduced by trade agreements, regulatory policy changes, or geopolitical events, which could have an impact on the whole SME financing landscape.

Another important factor in SME lending is liquidity risk, which represents the possibility that lenders won't be able to fulfil funding commitments, sell assets to pay for loan disbursements, or honour depositor withdrawal requests. Mismatches in the maturity profiles of assets and liabilities, reliance on short-term funding sources, or interruptions in the capital markets that restrict access to liquidity can all pose problems for SME lenders

when it comes to liquidity. Furthermore, unfavourable credit occurrences or market trends could lead to liquidity crises, forcing lenders to reduce lending, tighten credit requirements, or apply for emergency funding from government organizations or central banks in order to preserve stability and solvency. Given the changing regulatory environment and heightened scrutiny from regulatory bodies over consumer protection, anti-money laundering (AML), and know-your-customer (KYC) requirements, compliance risk is a major worry for small and medium-sized lenders. Investor confidence and public confidence in the integrity and stability of the financial system can be undermined by financial institutions' non-compliance with regulatory duties, which can lead to fines, penalties, legal liabilities, and reputational harm. To reduce regulatory risks and guarantee compliance with relevant laws and regulations, SME lenders need to manage complicated regulatory frameworks, keep up with regulatory developments, and have strong compliance processes in place.

Protecting the interests of lenders, borrowers, and other stakeholders in the financial ecosystem requires careful analysis and proactive risk management measures, which are inherent in the varied array of risks associated with SME financing. In an increasingly uncertain and dynamic business environment, SME lenders can reduce risks, improve financial stability, and support the growth and resilience of small businesses by implementing responsible credit risk assessment practices, improving operational resilience, and fortifying compliance frameworks. Furthermore, in order to address systemic risks, encourage responsible lending practices, and promote inclusive economic development that benefits SMEs and society at large, it is imperative that industry players, legislators, and regulatory agencies collaborate and share expertise.

### **1.1.3 Challenges associated with SME Lending**

The difficulties in financing to small and medium-sized enterprises (SME) in the US and India markets are covered in detail in this section. SMEs are essential to a region's overall economic growth, employment generation, and innovation. Fintech institutions have both possibilities and challenges when financing to these enterprises. SMEs frequently lack financial information, which makes it challenging for lenders to fairly evaluate their creditworthiness. Fintech faces challenges in obtaining dependable and consistent data about the SME community because of dispersed sources and different reporting standards. The financial reporting procedures used by SMEs are frequently manual, which causes errors, inefficiencies, and delays.

However, SMEs encounter opacity in the mechanisms used for fintech loan evaluation. The criteria for review, documentation requirements, and loan vary throughout financial institutions. This discrepancy makes the loan procedure much more difficult. Because of some tender processes, a lack of automation, and bespoke solutions, supplying SMEs can come at a disproportionately high cost. Agility and responsiveness are hampered by the widespread use of antiquated legacy systems by financial organizations. SMEs are expecting more and more seamless digital experiences. Fintech needs to make investments in mobile apps, internet platforms, and user-friendly interfaces. By utilizing state-of-the-art modelling tools along with forward-looking data sources (such transaction data, internet sales, and other alternate data), one can improve the SMEs' credit risk assessment. Risks inherent in SME loans include volatility in the economy, demographics, industry-specific issues, and political instability in the local area.

Fintech needs to enhance its operational models in order to produce sustainable returns, even though SME lending might be profitable. Regulations may have an impact on SME loan procedures. Serving SMEs while maintaining capital adequacy levels can be difficult, particularly for smaller financial institutions. Due to their risk aversion and strict credit

checks, SMEs frequently have difficulty obtaining bank financing. Fintech is bridging the gap by providing SMEs with creative loan options.

The landscape of SMEs lending is undoubtedly dynamic, and Fintech must change to satisfy these companies' changing needs. Financial institutions may efficiently manage risks and exploit the great potential of SME lending by adopting technology, reducing procedures, and cultivating strong partnerships. Financial transactions are only one aspect of supporting SMEs; other goals include promoting economic expansion and giving entrepreneurs the tools they need to effect positive change.

## **1.2 Research Problem**

The majority of fintech companies have been working to digitize the SME lending process in order to achieve state-through processing, which will improve decision-making, speed up disbursements, and reduce costs. Therefore, in order to effectively make risk-based decisions, a strong system for hedging credit risk is required. Alternate data is utilized to back this, ensuring that decision-making is more trustworthy and that reliance on credit scores from credit bureaus is kept to a minimum. Analyzing bank statement data as a substitute source of information for creating cash flow variables to aid in credit decision-making is one way to find a solution. Significantly little work has been done in the past on the subject of analyzing bank statements to produce insightful data that improves credit decision-making.

Previous study (Mehta, 2020) has analyzed the bank statements data to produce the bank statement analyzer to boost the credit decisioning process. However, the accuracy of classification of bank statement transactions was not adequate and this study aims at improving the accuracy of the classification of banking transactions to multiple cash flow classes and at the same time to explore the model for India Market to see if this could work with minimal adjustments. In addition to this, current study would also explore the novel

AI technology i.e. Large Language Models for classification of the bank statement transactions.

### **1.3 Purpose of Research**

As highlighted in above point, the purpose of this study is to improve a bank statement analyser for SME Lending developed by (Mehta, 2020) that could produce more accurate insights to support robust credit decisioning. The data used for this study comprises of 1,00,000 transactions for 3,300 small and medium enterprises in the US. Applying this model to 20,000 transactions pertaining to 450 SMEs in India Market to see if this model could work with minimal adjustments and could be generalized for different geographies. This quantity of transactions is sufficient to establish a proof of concept and valuable insights that could be replicated for any future number of transactions. The US Lending and India Lending Markets are the geographic areas that this study takes into consideration.

In addition to above, study would also explore the novel AI technology like Large Language Models for the classification use case of this study to segregate the bank statement transactions into the cash flow classes in order to generate the cash flow variables.

### **1.4 Significance of the Study**

Risk reduction in SME (small and medium-sized firm) lending is a crucial aspect of ethical financial behavior that is required to sustain economic growth, promote entrepreneurship, and preserve the stability of financial institutions. Small and medium-sized business loans require risk mitigation from a variety of perspectives, including social, structural, and economical aspects. Thus, a comprehensive understanding of the importance of risk reduction in SME lending requires an examination of its diverse impacts and implications.

Fundamentally, SME lending acts as a stimulant for economic growth by giving small enterprises access to capital, which is essential for fostering innovation, wealth creation,

and the creation of jobs. SMEs make up a sizable share of the world's economic activity and support social mobility, job creation, and GDP expansion. However, because of their small size, lack of collateral, and perceived higher risk, SMEs frequently encounter difficulties obtaining funding. Therefore, reducing risk in SME lending is crucial to guaranteeing that small companies can obtain the funding they require to maintain and grow their operations, make investments in innovation, and seize growth prospects.

The promotion of entrepreneurship and innovation is one of the main reasons why risk mitigation in SME finance is essential. SMEs are renowned for their adaptability, inventiveness, and capacity to create new goods, services, and business models that stimulate economic growth. However, SMEs' capacity to obtain inexpensive finance and successfully manage risks is what determines their level of success. Financial institutions empower entrepreneurs to pursue their business initiatives with confidence by reducing risk associated with SME loans. This allows them to have access to the necessary financial resources and support to overcome obstacles, capitalize on opportunities, and expand their enterprises.

Furthermore, by lowering the possibility of credit losses, loan defaults, and systemic disturbances within the financial system, risk mitigation in SME lending promotes financial stability. Due to things like information asymmetry, short credit histories, and greater volatility than larger firms, financing to SMEs is inherently risky. Ineffective risk mitigation can have a negative impact on lenders, increasing the risk of financial losses, liquidity issues, and deterioration of capital adequacy. Financial institutions can improve the resilience and stability of the banking industry and protect depositor funds by identifying, evaluating, and mitigating risks related to SME lending through the implementation of rigorous risk management techniques.

Additionally, reducing risk in SME lending encourages financial inclusion by giving credit to underrepresented groups such as women-owned firms, rural businesses, and



microenterprises. In order to increase livelihoods, create assets, and engage more fully in the formal economy, people and communities need to have access to finance. This is a crucial component of economic empowerment and social inclusion. However, because of things like little collateral, unofficial corporate structures, and remote locations, typical financing methods frequently leave these vulnerable groups out. Financial institutions can close the gap between supply and demand for credit by implementing risk mitigation measures that are suited to the requirements and limitations of underserved borrowers. This will encourage inclusive growth and lessen inequality.

In addition, lowering the chance of credit shortages, financial crises, and systemic disruptions—all of which can have significant repercussions for the economy as a whole—contributes to general macroeconomic stability. SMEs are vulnerable to economic shocks and externalities because they are essential to the operation of job markets, supply chains, and consumer spending patterns. Financial institutions lessen systemic risks, assist SMEs become resilient in the face of unfavourable occurrences, and lessen the impact of economic downturns by ensuring that they have access to stable and dependable sources of financing.

Almost all the finance industries including banks, Fintech, and NBFCs are aiming at digitalization. One of the most important functions all of them perform is Lending. Digital Lending would bring state through processing for the end-to-end process that would bring numerous benefits including cost-effectiveness, faster disbursements, auto decisioning and effective risk assessments. There is almost a new start-up every year who wants to lend money to SME Businesses by taking a reasonable risk and providing short term disbursements at a higher interest rate. Various alternate data sources are being used by these fintech industries to generate useful insights and measure risk of lending to consumers. They are targeting majority of those consumers who are rejected by the Banking institutions because they have undesired credit scores. Customers are approaching

fintech's because of their ability of faster disbursement of loans with minimal checks and prerequisites.

This is an extremely challenging field as well because lending money to dishonest and dangerous people could cause the industry to collapse financially and force it into bankruptcy. Therefore, before any money is provided to customers, a thorough risk evaluation is required. Transactional level data from the customer's payment history is one of the alternative data sources. This has the amazing potential to offer insightful information about the client without relying on his credit score or any other social media dataset that could contain noise of any type. There are several data science tools available that may be used to analyze banking transaction data and produce incredibly insightful findings. This topic serves as the foundation for my research because it has the potential to bring in business for the financial industry.

Our study is limited to SME financing because most members of this group lack the appropriate credit score and are heavily focused on different start-ups that are profitable from the loans they receive. In addition, compared to other portfolios, this group's loan demand is consistently strong. The term "cash flow" is crucial in finance since it describes the amount of money coming in and leaving out, and efficient corporate operations require maintaining a certain balance of cash. The goal of this research is to automate these cash flows with the use of artificial intelligence tools and technologies, producing important metrics that may expedite the loan approval process for customers. This study offers a rare chance to demonstrate the informational strength of bank statements as stand-alone sources and their potential to produce robust indicators for loan decisioning by automatic or assisted means.

In order to improve credit decisioning, transaction-level data that was taken out of bank statements is analyzed in this study. This study demonstrates how the examination of transaction-level data can yield a number of significant financial output variables,

including information about consumer debt capacity and lending risk associated with unsecured loans.

The following critical metrics were generated. Revenue is the first cash flow variable to be generated; it is the money from sales or from the investment of capital or assets before any costs, taxes, etc. are subtracted. The next class is expenditure, which includes money spent on paying employees, making payroll, and buying goods and services, among other things. The gap between revenue and expenses is known as operating profit. Operational income is another name for operating profit. This represents the total revenue of the company after all outlays have been deducted. The next variable this analysis produces is the Debt Service Coverage Ratio, or DSCR. The amount of cash flow available to settle outstanding debt is indicated by the DSCR. An alternative definition of this would be the Operating Profit to total Debt Service ratio. The entire amount paid toward principal and interest on a loan is known as debt service. Debt Repayment is the total amount that the business has paid back. Non-sufficient funds, or NSF, is a variable whose count indicates when a customer's account does not have sufficient funds in it. The acronym for debt-to-income ratio is DTI. The previously mentioned variables enhance the credit decision-making process for unsecured loans by explaining the SME's residual debt capacity. This cash flow study was produced using the most reliable data sources, which are bank statements in PDF format. Because there is little possibility of noise in the data, the result drawn from cash flow variables are highly reliable.

Trend analysis of cash flow variables, such as operating profits and revenue, is particularly important since it shows whether or not a company's performance over time is consistent. This is closely tied to their ability to return loans. The industry credit risk models and the previously described cash flow variables could be used to strengthen the models and produce results that are more reliable and more verifiable.

## 1.5 Research Purpose and Questions

The purpose of this study is to improve the bank statement analyser produced in the study (Mehta, 2020) for SME Lending that could produce meaningful insights to support robust credit decisioning to reduce the risk in lending to SME population. At the same time, study (Mehta, 2020) was confined to the United States market, current study also take data pertaining to India as well to find the applicability of the Bank statement analyser for India market. Large Language Models are also explored for the transaction's classification. A more robust and superior decision-making process would result in less risk. The quantity of transactions required ought to be sufficient to yield a proof of concept and to yield valuable insights that could be replicated for any number of transactions in the future. The US and India Lending Markets are the geographical areas that this study takes into consideration.

This research study has the following objectives.

- To enhance the list of keywords used by the study (Mehta, 2020) for categorizing banking transactions of small and medium enterprises in the United States.
- To enhance the performance of the machine learning model, build in the study (Mehta, 2020) for categorizing banking transactions of small and medium enterprises in the United States.
- To perform the classification of banking transactions of SME into the above identified categories using novel AI that includes Large Language Models. Comparing the results of traditional AI (ML model) and LLM classification for the banking transactions.
- To apply the above study for different geographical location i.e. India lending Market to see if the same model has worked well for another location so that it could be generalized.

## CHAPTER II: REVIEW OF LITERATURE

### **2.1 Introduction**

Although bank statements are widely acknowledged for providing extremely accurate cash flow data, they were hardly ever utilized in the past for SME lending auto decisioning. The primary cause of this is because complicated feature and trend computations demand a unique set of skills and subject expertise specific to the finance services industry. This document makes reference to a number of conference papers, journals, and dissertations in order to develop an understanding of similar or related studies that have been conducted in the past, documented their significant findings, and identified their shortcomings. In order to develop an understanding of the methods that will be employed in this study and to choose which should be used for the present topic of the study, a variety of literary works pertaining to those methods have been studied. In order to emphasize the importance of this study, other reliable websites with a wealth of relevant knowledge are also mentioned. These websites give a general idea of why this study is necessary in the present Fintech revolution age to support SME lending's digitization. Fintech start-ups in India and around the world rely heavily on digital SME lending for their revenue, and as they compete to take share of the leftover market—that is, the portion of the population that is not qualified for loans from banks—they are revolutionizing the digitalization of the lending process and driving significant technological advancements.

### **2.2 Digitalization in SME Lending**

Patel and Gupta (2021) examined how start-ups and digital transformation intersect, with a particular emphasis on the SME financing industry. Although this field of study is still in its infancy, it is gaining ground quickly. The authors' goal is to give a summary of how startups are undergoing digital transformation. The study used articles from the Dimension database to assess trends in digital transformation in start-ups from 2016 to 2022. Even if digital transformation is becoming increasingly popular, more research that is especially

targeted at start-up digital transformation is still needed. These results can help entrepreneurs who are getting ready to start up or run new businesses. The study acts as a broad roadmap for further research in this area.

Ghosh *et al.* (2019) had given a study on Fintech Lending in India Market. According to this research, India has the highest global Fintech acceptance rate of 87% and is home to the second-largest Fintech start-up base. According to this survey, the three Fintech categories of wealth management, digital payments, and digital loans are where the majority of start-ups are located in India. This work has provided several significant discoveries, which serve as the foundation for my current investigation. SME, Consumer, and Online Lending are the three main Fintech Lending categories, according to this study. These Fintech companies began to emerge in 2013 in an attempt to dominate the credit market. According to this paper, SME credit demand would exceed \$1 trillion USD by 2023, suggesting that this is a topic worth more investigation and that it will help to digitize the global lending market. This study also showed that Fintech companies were able to create their own credit scores from a variety of data sources because of the availability of Data Analytics and associated tools and technology. Large tech businesses, NBFCs, and start-ups are in intense competition with one another to gain a larger market share. Although they have comparatively less access to funding, start-ups have demonstrated great data analytics capabilities.

Major Fintech businesses have recently begun to enter the loan market as part of their expansion strategy. These companies include Google Pay, Ola, Amazon, True Caller, and others. The ability to generate significant customer data insights, acquire higher margins than other Fintech business models like payments and financial services, and take advantage of the enormous untapped market for unsecured loans are the main motivators for entering the loan market. Key success elements in this SME Lending include the presence of strong data analytics skills, successful relationships with other businesses, and low default rates as a result of highly calibrated models.

More findings are revealed by (Ghosh *et al.*, 2019) in their NASSCOM Report. Consumer credit demand is expected to rise in the upcoming years, and in 2020, data analytics will receive greater attention in order to improve credit modeling. According to this study, old lending methods required more time-consuming, physical background checks that weren't based on analytics. To add to that, unsecured lending has become the primary priority in order to assist the 300 million unbanked households that are in need of credit access. Historically, collateral-based lending was preferred. By 2023, the unmet loan gap for SMEs is predicted to reach USD 900 billion. This study suggests that transaction level data could assist businesses in comprehending client wants and providing them with loans that are specifically tailored to their needs. Additionally, this data is reliable and clean. The tools needed to analyse enormous amounts of transaction-level data and produce insights are now available because of recent developments in data science and big data. Additionally, this research has indicated that default rates may be significantly decreased by cross-referencing various data points, one of which is the cash flow data that is the subject of the current study.

Article (Bhakta, 2019) published in the Economic Times has reported that a large number of start-ups are engaged in the unsafe lending business. It detailed how a small baker was turned down for a loan by banks but was given money by a payment service provider who examined his payment transaction history before deciding to approve the loan. Organizations that handle these transactions have a pulse on company trends due to the quick digitization of payments. They can progressively grow their business and provide small loans to clients by analyzing the cash flows. Fintech startups aim to provide small working capital loans with short repayment terms. As a result, risks are reduced and credit behavior is monitored.

The paper (Cornelli *et al.*, 2023) claims that the rise of actors other than traditional banks is driving a significant global transformation in the credit landscape. Peer-to-peer and marketplace lending are two examples of the digital lending methods that have been more

and more popular in the last ten years among these disruptors. Large technological firms, or "big techs," have concurrently joined the credit sector and started offering what's known as "big tech credit." This study explores the information, motivations, and possible effects of these nontraditional credit options. The credit environment is still dominated by traditional lenders like banks and credit unions. But the sector is changing as a result of big tech credit and the emergence of digital lending arrangements. These new intermediaries challenge conventional wisdom by providing creative methods of providing credit. To understand the scale and dynamics of fintech and big tech credit, the authors assembled comprehensive data covering 79 countries from 2013 to 2019. This database provides a valuable resource for researchers, policymakers, and practitioners seeking insights into the evolving credit ecosystem.

The volume of fintech and big tech credit combined, or "total alternative credit," approached USD 800 billion worldwide in 2019. In places like Asia (which includes China, Japan, Korea, and Southeast Asia), Africa, and Latin America, big tech credit grew quickly. Due to regulatory reforms in China, fintech lending volumes decreased in 2018–19, although they continued to grow elsewhere.

The uptake of fintech and big tech credit is impacted by several factors. These alternative credit models are more likely to be adopted by countries with higher GDP per capita. Large tech and fintech lending are growing more rapidly due to higher markups in the banking sector. Bank laws that are less onerous promote their growth. The number of bank branches per capita is connected with higher adoption. Corporate settings that facilitate business are conducive to innovation. Well-developed legal structures make alternative credit possible. Diminished bank credit-to-deposit ratios indicate room to grow for alternative credit. The maturity of the bond and equity markets has an impact on the penetration of fintech and big tech lending.



Fintech and big tech credit appear to enhance traditional credit rather than replace it. These other formats improve credit accessibility in general, particularly for marginalized populations. Fintech and big tech credit represent a significant shift in the way credit is extended. The Covid-19 pandemic's effects and other current events will influence the direction of these new credit models. Big tech and fintech credit are so challenging established banking practices, offering a range of credit options, and upending the status quo. In order to comprehend the importance of these patterns for stability and financial inclusion, scholars and decision-makers must closely monitor these advancements.

Studies stated above (Ghosh *et al.*, 2019; Bhakta, 2019) have clearly indicated that there is a future scope in the digitalization of the SME Lending and major role would be played by the analytics in this digitalization journey in coming years. A previous study (Yadav and Thakur, 2018) also highlighted the rise in the lending businesses in India. Thus, the current research topic is really useful for the lending businesses across the globe.

By combining or incorporating the transaction data from bank statements into credit risk models, this project will produce insights that may aid in the digitalization of lending to small and medium-sized businesses. According to a NASSCOM study by Ghosh *et al.* (2019), there may be a complete revolution in lending by using data analytics in 2020. This study is one of the options for streamlining customer journeys by digitizing the auto decisioning part for quicker approval of loan requests from Fintech start-ups.

Futures FinTech (2019) claims that recent technological advancements have made it feasible to evaluate credit for small and medium-sized businesses very instantly, resulting in efficient lending. According to this article, while a fully automated approach would not be practical for larger loans, state-through-processing can be achieved to some extent via AI and data science. This will facilitate effective credit monitoring and quicken the credit decisioning process. It has provided numerous instances of how sentiment analysis and clustering helped in gaining understanding of the behaviors of customers.

According to an article presented in (Futures FinTech., 2019), bank statement analysis for SME lending research is implementing machine learning solutions for SME lending because it is possible to perform complex credit decisioning analysis using AI and ML to target large populations that need capital for their business needs. According to (Futures FinTech., 2019), bank statement analysis is an illustration of a better use of external data to provide better credit outcomes leading to data-driven credit evaluation. This study demonstrates the applicability of our research.

According to the study (E&Y, 2018), 51% of the UK's private sector revenue is provided by SMEs, with a turnover of £1.9 trillion. This market is expanding, and it demands products that are better suited to their requirements. This study indicates that prior to the 2007–2008 financial crisis, traditional banks and experts provided the majority of SME lending services; since the crisis, a number of new players, including Fintech and challenger banks, have arisen. As a result, the SME loan market is revolutionizing, and my research would be of great help in data-driven credit evaluation. The study also discussed the possibility of connecting and exchanging data with any partner or third party in the future using API-driven architecture. Given this information, research on Bank Statement Analysis might be turned into an API-based product that Fintech or financial institutions could use to enhance digital SME lending by giving data-driven credit evaluation.

According to the study (Yadav and Thakur, 2018), there was a 790.93% rise in personal loan requests in the US between 2012 and 2016. This demonstrates unequivocally that the number of loan requests is always increasing and that financial institutions require improved technologies and instruments to evaluate credit risk. Anne et al. (2014) mentioned that small and micro businesses face significant obstacles when trying to obtain credit facilities; nonetheless, in a nation such as Kenya, these businesses contribute to around half of all jobs.

Small and medium-sized enterprises (SMEs) are critical to the growth and development of jobs in the global economy. Access to funding remains a significant challenge for small enterprises, particularly in underdeveloped countries where traditional lending channels are often unreliable and unavailable. The digitalization of lending processes is one innovative way to address these problems; it raises SMEs' access to financing and reduces many of the risks associated with lending.

In order to improve credit evaluation, expedite the loan procedure, and boost overall productivity, digital SME lending integrates technology, data analytics, and cutting-edge financial platforms. Leading this digital revolution have been fintech companies, who have transformed traditional lending procedures by utilizing advances in artificial intelligence, machine learning, and data science. Fintech lenders are able to evaluate SMEs' creditworthiness more quickly and correctly than traditional banks by utilizing automated algorithms and alternate data sources.

The process of digitalizing SME lending entails the amalgamation of cutting-edge financial platforms, data analytics, and technology to optimize credit evaluation, optimize loan procedures, and augment overall efficacy. Using developments in artificial intelligence, machine learning, and data science to disrupt traditional lending methods, fintech companies have been at the forefront of this digital revolution. Fintech lenders are more competent than traditional banks to evaluate the creditworthiness of SMEs since they use automated algorithms and other data sources.

Additionally, technology has greatly accelerated the loan approval procedure, making it possible for SMEs to obtain funding in a timely and effective manner. Large amounts of paperwork, labour-intensive underwriting procedures, and protracted approval periods are hallmarks of traditional loan applications, which provide formidable obstacles to SMEs in need of quick funding. On the other hand, digital lending platforms provide SMEs with a more efficient application process, frequently enabling them to apply online and get

funding in a matter of days or even hours. In addition to making funding more convenient and accessible, this quicker turnaround time helps SMEs take advantage of growth possibilities and overcome obstacles related to money.

Digitalization has not only made financing more accessible, but it has also significantly improved SME lending risk management procedures. Fintech lenders are becoming more adept at identifying and reducing many kinds of risks, such as fraud, operational, and credit risk, by utilizing predictive modelling and sophisticated analytics. With the use of machine learning algorithms, lenders may proactively manage risk exposures and maximize portfolio performance by analysing enormous volumes of data to find patterns and anomalies suggestive of possible credit defaults or fraudulent activity. Moreover, risk-based pricing algorithms and strong risk assessment frameworks are frequently included in digital lending platforms, guaranteeing that loan terms correspond to each borrower's unique risk profile.

It's crucial to recognize, though, that digitalization in SME lending also brings with it new dangers and issues that must be managed. For example, the use of automated decision-making methods and non-traditional data sources may make issues with algorithmic transparency, bias, and data privacy worse. Furthermore, regulatory concerns about consumer protection, fair lending practices, and systemic risks have been highlighted by the Fintech lenders' and online lending platforms' explosive growth. Therefore, the responsibility for creating suitable regulatory frameworks and supervisory procedures to protect the integrity and stability of the financial system and encourage competition and innovation in SME lending is on legislators and regulators.

All things considered, the digitization of SME lending has the ability to completely change the small business finance scene by providing better risk management tools, increased lending process efficiency and transparency, and increased access to capital. But in order to fully reap these rewards, legislators, regulators, financial institutions, and

Fintech firms must work together and coordinate their efforts to promote an environment that is supportive of digital innovation and to handle new risks and challenges in the quickly changing field of SME financing.

### **2.3 Understanding Cash Flow analysis**

The study conducted by (Raju and Reddy, 2023) examines the financial performance of the State Bank of India (SBI) from 2014 to 2019. Examining the financial health of the biggest and most stable bank in India is the goal, and it will be done by using a wide range of financial methods, including ratio, trend, comparative, and common size analysis. The credit environment is still dominated by traditional lenders like banks and credit unions. But the sector is changing as a result of big tech credit and the emergence of digital lending arrangements. These new intermediaries challenge conventional wisdom by providing creative methods of providing credit. The authors gathered detailed data from 2013 to 2019 that covered 79 nations in order to comprehend the scope and dynamics of fintech and big tech credit.

The volume of fintech and big tech credit combined, or "total alternative credit," approached USD 800 billion worldwide in 2019. In places like Asia (which includes China, Japan, Korea, and Southeast Asia), Africa, and Latin America, big tech credit grew quickly. Due to regulatory reforms in China, fintech lending volumes decreased in 2018–19, although they continued to grow elsewhere. Fintech and big tech credit adoption is influenced by a number of reasons.

Higher GDP per capita nations are more likely to adopt these alternative credit models. Greater markups in the banking industry promote the growth of large tech and fintech lending. Less restrictive banking laws encourage their expansion. Higher adoption is correlated with fewer bank branches per capita. Innovation is encouraged by corporate environments that are easy to do business in. Legal frameworks that are well-developed facilitate alternative credit. Reduced bank credit-to-deposit ratios suggest potential for

expanding alternative credit. Fintech and big tech credit uptake is also influenced by the maturity of the bond and equity markets. Fintech and big tech credit appear to enhance traditional credit rather than replace it. These other formats improve credit accessibility in general, particularly for marginalized populations.

As a result, SBI's financial statements show a strong organization thriving in a changing environment. The bank's future course will be decided by its strategic emphasis on client experience, risk management, and digitalization. Thus, SBI's financial trajectory from 2014 to 2019 demonstrates both flexibility and stability, solidifying its position as a major participant in the Indian banking industry.

Fiengenbaum and Thomas (1986) noted the significance of cash flow analysis in comprehending a company's financial state. This study suggests that helpful information that could enable the proper decisions with the least amount of risk can be generated using forecasting and specific methodologies.

As a result of the findings, there was substantial support for automating cash flow processes in order to generate insights and provide small and medium-sized businesses with access to the credit market. According to statistics from a number of studies, including (Ghosh et al., 2019; Fiengenbaum and Thomas, 1986; Bhakta, 2019; Wadesango et al., 2019), a large number of small and medium-sized businesses are unable to get loans from financial institutions, which poses a threat to their ability to survive because they fail to implement proper cash management procedures. Fintech startups therefore aim to target these clients in order to generate income by offering them loans at a fair risk.

Poongavanam (2017) asserts that it is evident that cash flows can be utilized to predict profits, and this data can be utilized for additional study. As this journal has indicated, cash flows encompass a variety of information pertaining to operating, financing, and investment operations.

According to studies (Kahnamouei, 2017; Lokmic and Smith, 2000), operating cash flow modeling essentially helps creditors and investors make decisions that minimize credit risk. In order to predict future cash flow for sample enterprises, (Kahnamouei, 2017) suggested a strategy that uses a multi-layer perceptron neural network and a radius-based function neural network. However, for cash flow forecasting, (Lokmic and Smith, 2000) had employed both supervised and unsupervised neural networks. Because cash flow analysis gives lenders a thorough understanding of a borrower's financial situation, liquidity, and ability to repay debt, it is essential to the improvement of credit decisioning procedures. It provides a comprehensive picture of the financial performance and sustainability of people, companies, or projects by evaluating the cash inflows and outflows over a given time period. Lenders can reduce credit risk, maximize portfolio management tactics, and make well-informed lending decisions by incorporating cash flow analysis into credit decisioning. This study examine the importance of cash flow analysis in credit decisioning, as well as its applications, methodology, and consequences for both lenders and borrowers, in this in-depth discussion.

Fundamentally, cash flow analysis involves systematically analysing the inflows and outflows of cash from financing, investing, and operating operations in order to evaluate an entity's capacity to raise capital and pay its debts. In order to find trends, patterns, and anomalies in cash flow dynamics, it entails evaluating financial statements such as income statements, balance sheets, and cash flow statements. Through the assessment of variables like income production, cost control, investment decisions, debt servicing responsibilities, and working capital administration, cash flow analysis offers lenders important information about the health, stability, and solvency of borrowers.

Compared to standard credit scoring models or financial ratios alone, cash flow analysis offers a more precise and trustworthy evaluation of a borrower's repayment capacity, which is one of the main ways it improves credit decisioning. Even though financial ratios and credit ratings are useful tools for assessing creditworthiness, they might not fully

account for the subtleties of a borrower's cash flow patterns, especially when it comes to people or companies with complicated financial structures, erratic income streams, or seasonal swings. Lenders can assess a borrower's capacity to produce enough cash to pay off debt, maintain operations, and weather bad economic times with the use of cash flow analysis, which improves the precision and dependability of loan choices.

Furthermore, cash flow analysis helps lenders spot early warning indicators of trouble in the financial system, like declining profitability, cash flow mismatches, or liquidity restrictions, which could indicate a higher chance of default or delinquency. Through the examination of historical trends in cash flow metrics, including operating cash flow, free cash flow, cash conversion cycle, and liquidity ratios, lenders can evaluate the stability and longevity of borrowers' financial positions, spot possible warning signs, and proactively reduce credit risk. Declining cash flows, rising debt, or ongoing shortages of liquidity, for instance, could be signs of operational difficulties or financial strain that call for additional investigation and risk-reduction measures.

Cash flow analysis makes scenario analysis and stress testing easier in order to assess how unfavourable events or downturns in the economy may affect borrowers' projected cash flows and ability to pay off debt. Lenders can evaluate the robustness of borrowers' cash flow models and stress-test their capacity to survive unfavourable shocks by modelling various scenarios, such as modifications to revenue assumptions, changes in interest rates, or volatility in the market. By using scenario analysis, lenders can assess how potential negative outcomes might affect key loan performance indicators like loan-to-value ratios (LTV), loan loss provisions, and debt service coverage ratios (DSCR). This information can be used to guide capital allocation, risk management, and credit decisions.

Cash flow analysis also improves portfolio management and asset-liability management (ALM) strategies through capital allocation optimization, risk diversification, and lending activities that are in line with risk appetite and strategic goals. Lenders can find



possibilities for portfolio optimization, risk mitigation, and performance enhancement by comparing cash flow parameters across various loan portfolios, industries, or geographic areas. Lenders might, for instance, provide preference to loans with high recovery rates, low default probabilities, and great cash flow generating potential while limiting their exposure to industries or borrowers with higher credit risk or volatility. Furthermore, cash flow analysis helps lenders manage liquidity risks, maximize funding sources, and maintain sufficient capital buffers by offering insights into funding requirements, interest rate sensitivity, and liquidity demands.

Furthermore, by allowing lenders to watch borrower financial performance, keep an eye on changes in cash flow dynamics, and identify early warning indicators of credit deterioration, cash flow analysis helps with credit monitoring and loan servicing operations. Lenders can identify possible credit issues, proactively monitor loan portfolios, and mitigate losses by installing powerful credit monitoring systems and dashboards that combine real-time cash flow data. Lenders may be prompted to interact with borrowers, renegotiate loan terms, or undertake workout strategies to address emerging credit issues by automated alerts and triggers, for instance, which can alert them to significant deviations from projected cash flow forecasts, missed debt service payments, or declining financial ratios.

Because cash flow analysis gives lenders thorough information on the financial stability, liquidity, and repayment capability of borrowers, it is essential for improving credit decision-making processes. Lenders can reduce credit risk, enhance portfolio management techniques, and make better lending decisions by analysing cash flow dynamics, trends, and projections. Additionally, lenders may efficiently monitor credit performance, stress-test loan portfolios, and spot early warning signals of financial disaster thanks to cash flow analysis. Lenders can improve their capacity to evaluate creditworthiness, control risk, and facilitate sustainable lending practices that foster

resilience and financial stability by incorporating cash flow analysis into credit decisioning procedures.

#### **2.4 Methods used in past for Text Classification or Bank transactions classification**

According to the study (Egesa, 2016), text classification and bank transaction classification are two distinct jobs that need labeling or classifying data based on specific criteria or characteristics. Both jobs involve machine learning techniques and algorithms, but there are some significant distinctions between them in terms of the types of data that are used, the methods for extracting features, and the domain-specific challenges.

Text classification, sometimes referred to as text categorization or document classification, is the process of classifying written texts into pre-established groups or labels according to their content. Numerous applications, including topic modelling, emotion analysis, spam detection, and document organization, frequently use this job. Text classification algorithms examine the words, phrases, and semantic patterns found in documents to find pertinent patterns and correlations that set certain groups or classes apart.

The classification of bank transactions, on the other hand, entails grouping financial transactions according to characteristics including the type of transaction, the merchant category, the amount, and the date. In order to make financial administration, budgeting, and analysis easier, bank transactions are automatically categorized into predetermined categories, such as income, expenses, transfers, bills, or investments. Algorithms for classifying bank transactions examine transaction data, such as amounts, timestamps, and descriptions, to find trends and connections that set distinct transaction categories apart.

The type of data is one important distinction between classifying bank transactions and text. Emails, social media posts, articles, documents, and other unstructured or semi-structured text are common types of textual data used in text categorization tasks. To extract significant characteristics and patterns for classification, text classification

algorithms must process and examine the textual content of documents. On the other hand, structured or semi-structured data, like transaction records kept in databases or spreadsheets, makes up bank transaction data in activities involving the classification of bank transactions. In order to classify bank transactions into predetermined categories, algorithms for doing so concentrate on extracting pertinent transaction information, such as dates, amounts, and descriptions.

The techniques utilized for feature extraction distinguish text categorization from that of financial transactions. Word embeddings, topic modelling, TF-IDF (Term Frequency-Inverse Document Frequency), bag-of-words, and other feature extraction techniques are frequently used in text classification to represent textual texts as numerical feature vectors that may be fed into machine learning algorithms. The goal of these feature extraction techniques is to extract textual data's semantic meaning, context, and word relationships. On the other hand, transaction attributes like quantities, timestamps, and descriptions are frequently used as input features for the classification of bank transactions. Transaction data is converted into numerical representations appropriate for classification algorithms using feature extraction techniques such date/time parsing, numerical normalization, and one-hot encoding.

Because of the structure of the data and the particular application context, classifying text and financial transactions may present distinct difficulties and considerations. To get accurate classification results in text classification, robust preprocessing, feature engineering, and model selection strategies are necessary to address obstacles such data sparsity, ambiguity, linguistic variability, and semantic ambiguity. On the other hand, the classification of bank transactions may encounter difficulties like data imbalance, transaction variability, and granularity of classification; therefore, to guarantee efficient classification performance, transaction attributes, class definitions, and algorithmic approaches must be carefully considered.

While utilizing machine learning algorithms for categorization tasks is similar between text classification and bank transaction classification, there are substantial differences between the two tasks in terms of data properties, feature extraction techniques, and domain-specific problems. While bank transaction classification involves classifying structured transaction data according to transaction attributes, text classification concentrates on evaluating unstructured textual data to classify documents based on their content. For text classification and bank transaction classification tasks, it is crucial to comprehend these distinctions in order to choose the right algorithms, preprocessing methods, and assessment metrics that will produce accurate and dependable classification results.

The study's (Mytnyk et al., 2023) main objective is to employ artificial intelligence (AI) to detect bank fraud. Due in part to the COVID-19 pandemic, bank fraud has become more commonplace in recent years. The transition to digital platforms and the growth of charity donations have given thieves plenty of possibilities to trick users. In light of this, the study investigates machine learning algorithms as useful instruments for identifying and evaluating fraudulent online banking activities. The research creates machine learning models with the express purpose of detecting fraudulent bank transactions. These algorithms use past data to identify trends linked to dishonest behavior. The study describes feature engineering, feature modification, and handling extremely unbalanced datasets as preprocessing methods for bank data. These actions improve fraud detection accuracy. The suggested methodology considerably increases the accuracy of detecting fraudulent transactions and is based on an artificial neural network. The study presents various algorithms' performances graphically. Notably, logistic regression has strong performance, with an output AUC value of almost 0.946. The AUC of 0.954 obtained with stacked generalization is considerably better.

The study's key findings include the following: good feature engineering improves model performance; finding the ideal balance between computing efficiency and detection accuracy is critical; and identifying financial fraud using AI algorithms is an urgent issue

in our digital world. In summary, this study emphasizes how crucial artificial intelligence is to protecting banking processes. Banks may proactively detect and stop fraudulent activity, safeguarding both financial institutions and their clients, by utilizing machine learning and preprocessing approaches.

The research (Mahalakshmi and Duraiswamy, 2012; Thangaraj, 2018) offered a range of supervised and unsupervised text classification techniques along with the benefits and drawbacks of each. This was a crucial step in learning how to classify text (i.e., transactions descriptions) using different machine learning techniques, as my research also focuses on transaction classification. A distinct issue with financial transactions, though, is that their data dictionaries and categorization schemes should take into account the varied keywords that appear in the description of the transactions.

The analysis of financial transactions was further developed by research from (Egesa, 2016), which employed Apache Spark's machine learning library (MLlib) on a dataset of over 2.4 million categorized transactions to classify the financial transactions. SVM has fared better in this study than other classification techniques. In addition, this study indicates that for the words in the description, Count Vectorizer was utilized instead of TF-IDF. Instead of stop words, N-grams—preferably quad-grams—are utilized. Similar to this work, my research will likewise use TF-IDF, n-grams, and count vectorizer to create keywords from the transaction description in order to further classify the test dataset. Although the focus of my work is features extraction, my research challenge also involves banking transaction clustering and categorization.

In real life, there would be an enormous number of transactions, and it would be pointless to classify every single one of them because some of the transactions would be similar. Even if we could initially identify similarities between transactions by using clustering to group similar transactions together, we would still only be able to classify the most frequent transactions from each cluster. By expediting the process of categorizing the

transaction groupings rather than the individual transactions, this would improve process efficiency. The study conducted by (Guidotti et al., 2017) also indicates this.

There are several ways to conduct clustering; a work by (Omran et al., 2007) presents a comprehensive review of the many clustering approaches. Hierarchical clustering, K means clustering, and additional clustering techniques were presented in this paper. Singh (2018) claims that when it comes to text classification, Recurrent Neural Networks performed better than both conventional and cutting-edge machine learning techniques. Numerous difficulties with the text classification process were highlighted in this study. It was mentioned that machine learning for the feature extraction procedure in the text classification preprocessing work might be separated, hence resolving the feature engineering issue. This is how our research would proceed, with keyword extraction maintained as a separate module for increased efficiency and to speed up the classification process.

Kowsari et al. (2019) discussed a variety of text classification techniques and how well they performed in this study at different levels (document, paragraph, sentence, and sub-sentence levels). In order to solve time-complexity issues, the overall procedures in this study were feature extraction, which will extract important features from the text that is already available. Dimensionality reduction techniques were employed in this study, and classifiers were then added on top for categorization. This paper is really helpful for my research since it gives me a general understanding of how transaction descriptions, which are nothing more than sentences found in texts on banking, should be categorized. But this research is generic in nature and has nothing to do with the extraction of financial features or the categorization of financial transactions. This study suggests that using machine learning to classify bank transactions is an effective method to automate the classification of financial transactions into several classes or categories according to their attributes. Machine learning is being used extensively in a number of financial activities, such as financial analysis, fraud detection, spending tracking, and budgeting. Financial

institutions, fintech companies, and people can improve decision-making processes, obtain insights into expenditure patterns, and streamline financial management activities by utilizing machine learning algorithms.

Classifying financial transactions into predetermined groups or categories according to particular criteria, like transaction type, merchant category, transaction amount, and date, is the basic process of classifying bank transactions. The goal is to automate the process of labelling transactions, which will make duties like financial reporting, spending classification, and budget management easier. A data-driven method for classifying bank transactions is provided by machine learning techniques, which enable algorithms to infer patterns and associations from past transaction data and forecast the occurrence of future, unseen transactions.

Supervised learning is one of the main approaches used in machine learning for the classification of bank transactions. Labelled training data, in which every transaction is connected to a predetermined class or category, is the source of knowledge for supervised learning algorithms. Neural networks, k-nearest neighbours (KNN), support vector machines (SVM), decision trees, random forests, and other supervised learning techniques are frequently used for classifying bank transactions. In order to find patterns and relationships that differentiate various transaction categories, these algorithms examine transaction attributes such as transaction descriptions, amounts, dates, and merchant names. Supervised learning algorithms are capable of accurately classifying transactions into predetermined classes based on their attributes by training on labelled data and modifying model parameters.

In order to enable algorithmic learning, feature engineering—which includes choosing, converting, and encoding transaction attributes—is essential to machine learning for the classification of bank transactions. Often, input features for classification algorithms are transaction attributes including dates, sums, merchant names, and descriptions of the

transactions. Transaction data is converted into numerical representations that are appropriate for machine learning algorithms using feature engineering techniques such as text tokenization, numerical normalization, one-hot encoding, and date/time parsing. To extract informative aspects from transaction data, such as recognizing keywords or patterns in transaction descriptions that point to particular transaction categories, domain-specific knowledge and heuristics may also be applied.

Additionally, in situations when labelled training data is few or non-existent, unsupervised learning methods like clustering and anomaly detection can be applied to classification tasks involving bank transactions. Without the use of predetermined class labels, unsupervised learning algorithms examine transaction data to find clusters or patterns that distinguish between various transaction categories. By grouping related transactions together based on shared characteristics, clustering methods like k-means clustering, hierarchical clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) make it possible to identify common spending patterns or transaction clusters. Similar to this, anomaly detection algorithms like autoencoders and isolation forests can spot odd or fraudulent transactions that diverge from typical spending patterns, aiding in the detection of possible fraud or anomalies in financial transactions.

Additionally, machine learning models for the classification of bank transactions can be made more robust and perform better at generalization by utilizing ensemble learning techniques like bagging, boosting, and stacking. In ensemble learning, several base classifiers are combined to produce a stronger, more accurate classifier that performs better than any one model alone. To better capture complex correlations and patterns in transaction data, ensemble approaches like gradient boosting machines (GBM) and random forests, for instance, combine several decision trees to construct an ensemble model. Financial institutions can increase the precision, dependability, and scalability of machine learning models for classifying bank transactions by utilizing ensemble learning



approaches. This will improve decision-making procedures and lower the possibility of misclassification mistakes.

Machine learning for the categorization of bank transactions presents a number of issues and concerns, including feature selection, class imbalance, data quality, interpretability of the model, and regulatory compliance. Problems with data quality, like missing values, duplicate entries, and inconsistent data, can impair machine learning model performance and produce unreliable classification outcomes. To guarantee that transaction data is correct, comprehensive, and dependable, thorough data pretreatment, cleaning, and validation processes are necessary in order to address data quality concerns. Furthermore, class imbalance might skew model predictions and result in less-than-ideal performance if specific transaction categories are underrepresented in the training set. Techniques like oversampling, under sampling, or employing class-weighted loss functions to modify the weight of minority classes during training can all be used to mitigate class imbalance.

Another important factor in machine learning for the classification of bank transactions is feature selection, which is the process of locating useful features that are crucial for differentiating between various transaction categories. The performance of classification might be hampered by noise or irrelevant information in transaction attributes including transaction amounts, merchant names, and transaction descriptions. Machine learning models can be made more efficient and effective by using feature selection techniques like dimensionality reduction, feature importance ranking, and recursive feature elimination (RFE) to find the most useful features and eliminate unnecessary or redundant ones.

Machine learning for the classification of bank transactions requires interpretable models, especially in regulated businesses where accountability and transparency are critical. It can be difficult to interpret complex machine learning models, such neural networks and ensemble techniques, which makes it difficult to comprehend the fundamental

mechanisms guiding model predictions. Interpretability strategies that aid in understanding machine learning model decision-making and identifying elements that influence categorization outcomes include feature importance analysis, partial dependence graphs, and model-agnostic explanation tools. Furthermore, the gathering, processing, and storage of transaction data are restricted by regulatory compliance considerations, such as data privacy laws (such as the GDPR, CCPA), and anti-money laundering (AML) mandates. To safeguard sensitive data and maintain regulatory compliance, these considerations call for strong data governance, security, and compliance measures.

A data-driven method for automating the classification of financial transactions into groups according to their characteristics is provided by machine learning for bank transaction classification. Techniques for supervised learning, unsupervised learning, and ensemble learning can be used to create precise and trustworthy classification models that speed up the detection of fraud, financial management, and decision-making. But in order to guarantee the efficiency and moral use of machine learning in the classification of bank transactions, problems with data quality, class imbalance, feature selection, model interpretability, and regulatory compliance need to be resolved. Financial institutions, fintech firms, and individuals can attain their financial objectives by optimizing financial operations, gaining knowledge about spending trends, and making well-informed judgments by utilizing machine learning algorithms and methodologies.

Sun et al. (2023) has mentioned in their study that there has been a notable surge in interest in large-scale Language Models (LLMs) such as GPT-3 in the past few years. They still perform worse than refined models in text categorization tasks, though. The limited number of tokens permitted for in-context learning and the inability to reason about intricate linguistic phenomena (such intensification, contrast, and irony) are two of the constraints. Researchers developed Clue and Reasoning Prompting (CARP) to overcome these problems. A progressive reasoning approach designed specifically for text classification is

used by CARP. It first asks LLMs to recognize surface-level cues (such as terms, tones, and semantic relationships) that direct a diagnostic reasoning procedure. Furthermore, CARP integrates optimized models with LLMs, taking advantage of both generalization capabilities and task-specific information from labelled datasets. On popular text-classification benchmarks such as SST-2, AGNews, R8, and R52, CARP impressively achieves state-of-the-art (SOTA) performance. Moreover, CARP shows impressive performance in settings involving low resources and domain adaptation.

Text classification using Large Language Models (LLMs) is a cutting-edge method for classifying textual data into predetermined classes or categories based on the semantic and contextual comprehension of language, claims the paper (Sun et al., 2023). By utilizing enormous volumes of pre-existing text data to discover intricate linguistic patterns and relationships, large language models—like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers)—have completely changed natural language processing tasks. More precise, contextually aware, and nuanced text data classification is made possible by LLMs and can be applied to a variety of tasks, such as sentiment analysis, document classification, spam detection, and automated customer service. We examine the methods, strategies, benefits, difficulties, and uses of text classification using Large Language in this in-depth conversation.

Essentially, large language models text classification is adjusting previously trained language models on labelled training data to customize them for particular text classification tasks. Unsupervised learning approaches are used to pre-train large language models, like GPT and BERT, on massive text corpora. This allows the models to acquire contextual representations of words, phrases, and sentences. These models gain a thorough comprehension of language semantics, syntax, and context during pre-training, which equips them to recognize intricate linguistic relationships and patterns. In order to fine-tune the language models, more training on task-specific labelled data is required. This allows the model to learn how to map input text sequences to appropriate class labels or categories.

The ability of large language models to capture contextual information and the semantic relationships between words and phrases is one of the main benefits of text classification using these models. Text is often represented as bag-of-words or TF-IDF vectors in traditional machine learning models, like Naive Bayes or Support Vector Machines, which may lose contextual and semantic information. Large Language Models, on the other hand, use self-attention techniques to encode text input as continuous vector representations, which allows them to capture syntactic structures, semantic subtleties, and long-range connections in text data. Large Language Models can comprehend word meaning in the context of the complete sentence or document by utilizing contextual embeddings, which makes text classification more precise and contextually aware.

Furthermore, transfer learning—the process of using pre-training knowledge from large-scale text data to subsequent text classification tasks—benefits text classification using big language models. Text input is encoded into dense, high-dimensional vector representations using pre-trained language models, such as GPT and BERT, which perform as potent feature extractors. Pre-trained language models are taught to adjust their representations to the unique subtleties and features of the target classification job by means of fine-tuning them on task-specific labelled data. Text classification models can get cutting-edge results with less labelled training data thanks to transfer learning, which allows them to apply pre-trained knowledge to generalize to various text classification tasks.

Additionally, more reliable handling of uncommon terms, domain-specific terminology, and out-of-vocabulary (OOV) words is made possible by text classification utilizing large language models. Large Language Models can learn representations of uncommon or domain-specific terms that might not be present in the training data since they are trained on enormous volumes of text data from various sources and domains. Furthermore, large language models are able to generalize to previously encountered words or phrases depending on their context by capturing contextual information and semantic similarities

between words. Consequently, large language model-based text classification models show better robustness and generalization capacity, especially in fields containing jargon or special terminology.

Large language models for text classification have many benefits, but in order to create trustworthy and efficient classification models, a number of issues and concerns need to be taken into account. The computational resources needed to train and optimize big language models on massive amounts of text data are one difficulty. Large quantities of memory and potent GPUs or TPUs are among the computational resources needed to train large-scale language models like GPT-3 or BERT. It takes a lot of time and computing power to fine-tune pre-trained language models on task-specific data, especially for big datasets or challenging classification tasks. To properly train and deploy Large Language Models, addressing these computational issues would necessitate having access to cloud-based services or high-performance computing equipment.

Also, transfer learning and domain adaptation strategies are crucial for customizing language models that have already been trained for certain text categorization tasks and domains. Pre-trained language models might not accurately reflect the subtleties and features of the target classification task or topic since they are usually trained on general-purpose text data from a variety of sources and domains. Pre-trained language models can adjust their representations to the unique linguistic patterns and task-specific terms by fine-tuning them using task-specific labelled data. Text classification models' performance and capacity for generalization on domain-specific tasks can be enhanced with the use of domain adaption approaches such adversarial training, multi-task learning, and domain-specific pre-training.

Furthermore, when classifying texts using large language models, interpretability and explainability of the model are important factors to take into mind, especially in applications or businesses that are subject to regulations and where accountability and

transparency are crucial. With millions or billions of parameters, large language models like GPT and BERT are intricate neural network designs that are intrinsically challenging to grasp and comprehend. It can be difficult to interpret how Large Language Models make decisions and to pinpoint the elements that influence categorization results, especially for stakeholders or non-experts who lack a thorough understanding of machine learning methods. Methods for identifying saliency maps, attention visualization, and feature attribution are examples of model interpretability strategies that assist stakeholders in comprehending how Large Language Models generate predictions and pinpoint significant features or tokens that impact categorization choices.

A cutting-edge method for classifying textual data based on semantic and contextual comprehension of language is text classification using large language models. Text classification models can provide higher performance, resilience, and generalization ability across multiple domains and applications by utilizing pre-trained language models, such as GPT and BERT. To create efficient and trustworthy text classification models, however, a number of issues including computing capacity, domain adaption, and model interpretability need to be resolved. Through the utilization of Large Language Models for text classification, organizations can automate decision-making procedures, gain valuable insights from textual data, and seize new opportunities for innovation and growth by surmounting these obstacles and capitalizing on advancements in natural language processing techniques.

## **2.5 Problem statement**

Nothing particular was discovered about the classification of bank statements transactions for SME financing, as mentioned in the section above. The terminology used in bank statements varies greatly depending on the bank, the kind of customer, the type of transaction, and the region. Therefore, it was not possible to employ generic text classification algorithms directly in this study.

Bank statement data is used as an alternate data source in my research, which is similar to this study (Ghosh et al., 2019), which shows that there might be a considerable reduction in default rates if different data can be used for creating meaning insights. According to a study (Futures FinTech, 2019), my research would use machine learning and analytics to do complicated analysis of bank statements.

Harnessing alternative data sources, such as bank transactions, presents a transformative opportunity to gain valuable insights into consumer behaviour, financial health, and creditworthiness. Traditional credit assessment methods often rely on limited data points, such as credit scores and income statements, which may not fully capture an individual's or a business's financial situation. In contrast, leveraging alternative data sources enables lenders, financial institutions, and businesses to access a richer and more comprehensive dataset, providing deeper insights and more accurate risk assessments. By analysing bank transactions, which encompass a wealth of information about spending patterns, cash flows, and financial habits, stakeholders can make more informed decisions, improve credit underwriting processes, and enhance financial inclusion.

Bank transactions provide a thorough record of income sources, spending trends, and saving behaviour, acting as a digital footprint of a person's or company's financial activity. Lenders are able to determine a person's creditworthiness more precisely by analysing these transactions, which can provide important insights on the financial behaviour, lifestyle choices, and risk propensity of the individual. In contrast to inconsistent revenue patterns, excessive spending, and overdrafts, which may suggest increased credit risk, frequent deposits from steady income sources, consistent bill payments, and careful spending habits may show financial stability and creditworthiness. Lenders can improve the accuracy and dependability of credit judgments by analysing transaction data at scale, seeing predicted trends, and creating credit scoring models that take into account more variables by utilizing machine learning algorithms and data analytics approaches.

Likewise, bank transactions offer lenders and other financial institutions real-time visibility into cash flows and liquidity levels, which improves their ability to monitor borrowers' financial stability and ability to repay debts. Lenders are better equipped to recognize changes in borrowers' financial conditions, foresee liquidity issues, and proactively interact with borrowers to provide support or adjust loan terms, when necessary, by monitoring borrowers' income inflows, spending patterns, and savings habits over time. Lenders may take early action and reduce credit risk in response to indications of financial trouble or imminent default, such as abrupt changes in income, erratic spending habits, or unforeseen expenses. Financial institutions can lower losses, improve portfolio performance, and decrease loan delinquencies by using transaction data for proactive credit management and dynamic risk monitoring.

Based on individual tastes and financial objectives, bank transaction analysis can support targeted marketing campaigns, customized product suggestions, and personalized financial advice. Businesses may provide customers with appropriate products and services that fulfil their demands and increase customer satisfaction by getting to know their spending patterns, lifestyle preferences, and financial priorities. For instance, based on clients' transactional activity and life events, banks might use transaction data to find potential for cross-selling other items like insurance policies, investment products, and savings accounts. Fintech startups and financial technology firms have the ability to utilize transaction data to create creative solutions that address specific pain areas and enable customers to make better financial decisions. Examples of these solutions include personalized loan offers, financial management tools, and budgeting apps.

Additionally, bank transactions offer insightful information about consumer preferences, economic indicators, and market trends. This information helps firms and governments recognize new opportunities, reduce risks, and make data-driven decisions. Researchers and analysts can evaluate economic activity at a specific level, analyse consumer purchasing trends, and perform macroeconomic analysis by combining and anonymizing



transaction data from several sources. For instance, examining transaction data from various industries, geographical areas, and demographic groups can reveal information about retail sales patterns, consumer confidence, and company sentiment. This information can help companies better plan for future changes in the market, modify their product offerings, and deploy their resources. In a similar vein, decision-makers in charge of monetary policy might utilize transaction data to guide their choices, evaluate the results of stimulus packages, and create focused interventions.

Deeper insights into market trends, financial dynamics, and consumer behaviour can be obtained by utilizing other data sources, such as bank transactions. Through the utilization of sophisticated analytics tools and machine learning algorithms to analyse transaction data, interested parties can optimize credit underwriting procedures, elevate risk management strategies, and open up novel business avenues. Moreover, companies and legislators may better serve customer requirements, stimulate innovation, and promote economic growth by utilizing transaction data for targeted marketing, individualized financial advice, and market information. To ensure that data-driven innovation helps consumers while protecting their rights and interests, it is crucial to address privacy concerns, data security threats, and ethical considerations related to the use of alternative data.

The research would employ the same methodologies as the study (Egesa, 2016) in order to generate keywords based on the transactions recorded in the bank statements. This study gave a clear summary of feature extraction, which would be used to the Bank Statement's keyword creation process.

It has also been stated, as well, that the classification of text differs from that of bank statement transactions due to the existence of specialized terminology used in the banking sector, which is heavily reliant on banks, locations, and clients. It has been noted how text classification differs from bank statement transaction classification, and this raises an

interesting research question. The digitization of SME lending has been shaped by fintech start-ups using various strategies. In my research, I have proposed that bank statement analysis may yield outcomes that would facilitate their credit decisioning procedures.

## 2.6 Gaps identified in the Literature Review

Following table elaborates the gaps that have been found in literature review process –

Table 1 Gaps identified in Literature Review

S.No.	Referred Paper	Citation	Summary	Relevance to our study (Gaps if any identified)
1	The Role of FinTech in Small Business Lending	(Beaumont <i>et al.</i> , 2022; Rao <i>et al.</i> , 2023)	These papers highlight that Small and Medium Enterprises have limited access to credit after 2008 financial crisis due to increased regulations over the banking institutions. However, the emergence of the FinTech's have improved the credit access for SMEs by providing the unsecured lending	This paper has clearly indicated the fact that gap that has been created due to restricted lending to SMEs has been bridged by the emergence of the Fintech firms and they have made quite good profit by lending unsecured loans.

			by alleviating the collateral based secured lending.	
2	The Rise of Finance Companies and FinTech Lenders in Small Business Lending	(Gopal and Schnabl, 2022; Goel and Rastogi, 2023)	According to this article there is a drop of around 20% in bank lending from 2008 financial crisis till 2010. Emergence of FinTech's has improved the lending and played crucial role to boost the economy during this period. In the year 2016, fintech's provided 56% new loans.	The technology also plays important role in lending and Fintech platforms should bring the innovation to provide lending quicker and effective. The objective of my study is to use the alternate data to bring the state through processing in the SME lending process. However, this referred study does not list what innovations and technical advancements could be used for the effective lending.
3	Challenges small and medium enterprises (SMEs) face in acquiring loans	(Nassoro and Jaraj, 2022)	This study highlights that SMEs face challenges due to high interest rates, shorter repayment	This study did not suggest the solution to overcome the challenges faced by the SMEs in Tanzania.

	from commercial banks in Tanzania		period and lack of collaterals in acquiring loans from the commercial banks.	
4	Does bank complexity during the COVID-19 crisis alter the financing mechanism for small and medium-sized enterprises?	(Chen <i>et al.</i> , 2022)	This study provides the insights related to the correlation between the risks and complexity for banking institutions, more complex the bank is better it would be to manage the risks. However strict regulators restrictions disable them to take significant risks like lending SME population as they don't have consistent flow of income and better credit scores.	How could we minimize the risks that could arise as a result of lending SME population is not explained in this study and due to the unavailability of confidential data, regressor used could not guarantee the results that are expected out of this study.

5	Expectations of access to debt finance for SMEs in times of uncertainty	(Calabrese <i>et al.</i> , 2022)	This article examines the impact of Brexit referendum on the access to debt finance pertaining to the SMEs in the UK. How the expectations of the SMEs to be financially constrained have been reduced as a result of the referendum has been explained elaboratively in this study.	This study is specific to the recent change pertaining to the Brexit referendum on SME financing that is not generic and moreover how to mitigate this impact is not clearly explained in the article.
6	Classification of E-commerce Financial Transaction Logs Using Machine	(Ajithkumar and Geetha, 2022)	This article examines e-commerce financial transactions logs to figure out if there are any suspicious	The dataset used contained 5000 transaction logs that is very limited where as my study would use substantial data for classification. Moreover, very limited

	Learning Approach		transactions using the machine learning techniques. Support Vector Machine and Logistic Regression were used for transactions classification.	techniques were defined for the classification of transactions. Nothing has been specified to classify transactions with multi class labels.
7	Machine learning for financial transaction classification across companies using character-level word embeddings of text fields	(Jørgensen and Igel, 2021)	This research has used word embeddings and bag of words representation. The objective of this study was to map the financial transfers against the corresponding accounts.	Our study is involving large banking financial services vocabulary based on which multi class classifier has to be build. Similar to this study word embeddings like count vectorization could be used in my study and something beyond has to be used as idea behind my study is to create multi words keywords for classification.
8	How Digital Banking has Brought Innovative	(Haralayya, 2021)	This article focuses on the importance and prioritization of the digitalization in	Loan Management was identified as one of the focus areas for the digitalization but nothing is specified how

	Products and Services to India		the financial institutions. What are the various areas where digitalization is possible for achieving smoother customer journeys.	that could be achieved.
9	FinTech: Ecosystem, Opportunities and Challenges in Saudi Arabia	(Albarrak and Alokley, 2021)	Fintech in Saudi Arabia have taken various initiatives that have brought innovation and played a significant role in the future financial system in the Saudi Arabia.	This study has not listed any specific innovation and technology that Fintech had used and could be utilized industry wide for scaling up the economy further.
10	SMEs with legally restricted banking access: evidence from the US marijuana industry	(Merz and Riepe, 2021)	According to this study small and medium enterprises have very limited access to Banking due to legal constraints. This study suggested the importance of	However, this study does not specify how the access to financial services could be adequate in the developed economies for the SME.

			availability of access to the financial services for SME in the developed economy.	
11	A Comparison of Pre-Trained Language Models for Multi-Class Text Classification in the Financial Domain	(Arslan <i>et al.</i> , 2021)	According to this study vocabulary of the dataset of the financial domain have significant effect on the performance of the models. Generic models for multi class text classification models perform better than the specific models like Fin BERT model.	Two datasets that are generic pertaining to financial domain have been used in this study. In my study, multi class classification problem is to be dealt but the finance vocabulary is extremely vast and comprehensive due to significant variations in formats and other factors related to the Bank statements.
12	Fintech and Big Tech Credit: What Explains the Rise of	(Cornelli <i>et al.</i> , 2021)	According to this study, Fintech firms and Big-tech firms have been providing	Fintech are growing rapidly around the world and there is a big potential for different



	Digital Lending?		more lending to the SME population. Around USD 223 billion credit flows are reported for Fintech in the year 2019.	types of lending to boost economic growths. Hence this area is quite interesting for research from the overall growth perspective with respect to credit flows.
13	Digital Banking Support to Small Businesses amid COVID-19	(Sun, 2021)	The COVID-19 pandemic has introduced various challenges in accessing bank financing for small and medium enterprises, as per this article. Credit risk has been inflated and need different treatment to overcome the aftermaths of the same. Greece, Italy and others have developed the digital platforms to	This article has highlighted the importance of digital transformation to overcome the challenges introduced by the pandemic in credit lending. Our research objective is to bring the state through processing by digitalization in the field of SME lending.

			overcome the credit access limitations as a result of the pandemic.	
14	A systematic literature review on SME financing: Trends and future directions	(Rao <i>et al.</i> , 2021)	As per this study, Small and Medium Enterprises are the source of employment for many people and major constraint they have is the limited access to the finance. This paper has included enormous amount of literature review including 280 papers from 1986 till 2020 to conclude the finance research for the SMEs.	This study consolidates the past researches to understand the SME financing but does not specify how to overcome the limited finance access challenge using specific methodology to enable credit access for the SME population.
15	Deep learning in finance and banking: A	(Huang <i>et al.</i> , 2020)	This study analyzed the limited literatures available	According to this study after analyzing the literature between 2014 till 2018,

	literature review and classification		specific to the application of the deep learning in banking financial services domain. Given the complexity associated with the deep learning models, it was a challenge to justify the findings as per most of the papers analyzed by this study. Another challenge highlighted was that training these models was highly time consuming.	random forest algorithm yielded best accuracy for classification problems related to delinquency where minority class is rare posing a challenge of imbalance data.
16	Digital technologies in lending small and medium-	(Semenyuta <i>et al.</i> , 2019)	This article outlines the important role of the digital technologies in the system of the SME	One method described in this article is the automation of the process of interaction with the borrower but has a limitation as technical design

	size enterprises in Russia		lending in Russia specifically. Fintech starts-ups operates on online business development model as compared to the traditional banking that employs classical methods.	or real implementation process is missing. Another important finding in this article is the possibility of issuing credit without documents, guarantee etc. but how could that achieved with a reasonable risk is not explained.
--	-------------------------------	--	---	--

## 2.7 Summary

Nothing particular was discovered about the categorization of bank statements transactions pertaining to SME lending. The terminology used in finance on bank statements varies greatly depending on the bank, the kind of customer, the type of transaction, and the region. Therefore, it was not possible to employ generic text classification algorithms directly in this study. Bank statement data is used as an alternate data source in my study, which is similar to this study (D. Ghosh, A. Gupta, 2019), which shows that there might be a considerable reduction in default rates if different data can be used for creating meaning insights. According to a study (Futures FinTech, 2019), my research has embraced the machine learning and analytics technique for carrying out complicated analysis of bank statements. Beyond that trying Large Language Models for the classification of the bank statements transactions is an important objective of my study. In addition to this, not limiting to US but expanding the research to the India Market is another important research question related to my study.

The research would employ the same methodologies as the study (Egesa, 2016) in order to generate keywords based on the transactions recorded in the bank statements. This study offered a useful summary of feature extraction, which would be applied to the Bank

Statement Analyzer's keyword creation process. It has also been indicated, as well, that the classification of text differs from that of bank statement transactions due to the existence of specialized terminology used in the banking sector, which is heavily reliant on banks, locations, and clients. It has been noted how text classification differs from bank statement transaction classification, and this raises an interesting research question. Fintech startups have implemented methods to influence the digital transformation of SME lending, and my research suggests that bank statement analysis could yield outcomes that strengthen their credit assessment procedures.

CHAPTER III:  
METHODOLOGY

**3.1 Overview of the Research Problem**

The majority of Fintech companies have been working to digitalize the SME lending process in order to achieve state-through processing, which will improve decision-making, speed up disbursements, and reduce costs. Therefore, in order to effectively make risk-based decisions, a strong system for hedging credit risk is required. Alternate data is utilized to support this, ensuring that decision-making is more trustworthy and that reliance on credit scores from credit bureaus is kept to a minimum. Analyzing bank statement data as a substitute source of information for creating cash flow variables to aid in credit decision-making is one way to find a solution. Significantly little work has been done in the past on the subject of analyzing bank statements to produce insightful data that improves credit decision-making.

Fintech companies can improve risk management procedures, streamline credit underwriting procedures, and increase underprivileged groups' access to financing by utilizing alternative data sources. Alternative data is being used by the fintech sector more and more to enhance conventional credit scoring models and improve the accuracy of creditworthiness assessments for both individuals and companies. Conventional credit scoring models frequently depend on sparse data items, such income, work status, and credit history, which might not adequately reflect a person's or a company's financial circumstances. Fintech companies may create more comprehensive credit risk models by incorporating aspects like social connections, savings habits, and spending patterns into their models thanks to alternative data sources, which provide a more comprehensive and all-encompassing perspective of a borrower's financial activity. Fintech companies can find creditworthy borrowers who were possibly missed by traditional lenders by combining the analysis of alternative data with traditional credit metrics. This expands

the pool of potential borrowers for financing, especially for underserved groups such as small businesses, freelancers, gig workers, thin-file or no-file borrowers.

Fintech companies are also utilizing alternative data to improve identity verification and fraud detection procedures, which lowers the frequency of fraudulent activity and boosts the security of financial transactions. Conventional techniques for verifying identity, like document verification and credit bureau checks, can be vulnerable to financial fraud, identity theft, and synthetic fraud. Fintech organizations may create stronger identity verification procedures that withstand the strategies of fraudsters by utilizing alternative data sources including biometric data, device fingerprints, and behavioural analytics. For instance, biometric authentication techniques, like fingerprint scanning and facial recognition, reduce the possibility of identity theft and allow safe, convenient access to financial services. Similarly, to identify unusual behaviour suggestive of fraudulent activity, including account takeover attempts, unauthorized transactions, or suspicious login attempts, behavioural analytics algorithms examine user interaction patterns, transaction history, and device attributes.

Fintech companies are also utilizing alternative data to enhance loan underwriting procedures, more precisely determine credit risk, and customize loan products to each borrower's unique requirements and preferences. Fintech companies can create customized credit offerings that cater to the needs and preferences of borrowers by gaining insights into their financial behaviour, lifestyle preferences, and risk profiles through the analysis of alternative data sources like transaction data, social media activity, and online shopping behaviour. Fintech lenders, for instance, can evaluate cash flow dynamics, liquidity positions, and repayment capacity using transaction data. This allows them to provide flexible terms for repayment, dynamic pricing, and loan structures that are tailored to the risk profiles and financial situations of their borrowers. Similar to this, social media and online shopping activity analysis can reveal information about the interests, preferences, and spending patterns of borrowers. This information helps fintech

lenders create customized product recommendations, cross-selling opportunities, and marketing campaigns that appeal to borrowers and increase customer engagement.

Plus, alternative data is being used by the fintech sector to improve risk management procedures and more successfully reduce credit risk. Fintech companies can monitor changes in borrowers' financial conditions, avert defaults or delinquencies, and uncover early warning indications of financial trouble by proactively intervening through the analysis of alternative data sources, such as public records, social media activity, and online reviews. Social media activity monitoring, for instance, may highlight changes in borrowers' spending patterns, lifestyle decisions, or work status that may have an effect on their capacity to repay debts. Comparably, studying online reviews and customer comments can reveal information about how satisfied customers are, how loyal they are to a brand, and how likely they are to make payments. This information helps fintech lenders spot possible credit risks and implement risk-reduction strategies like renegotiating loan terms, providing financial counselling, or starting collection efforts. Fintech businesses are also using artificial intelligence (AI) and machine learning algorithms to analyse vast amounts of alternative data and spot predictive patterns. This helps them spot fraud, improve risk management plans, and make more accurate credit decisions.

Fintech companies are using alternative data to enhance credit underwriting procedures, increase loan availability for marginalized groups, and better manage risks. Fintech companies can obtain valuable insights into consumers' financial behaviour, lifestyle preferences, and risk profiles by utilizing alternative data sources, including transaction data, social media activity, and online behaviour. This information can help them create customized credit offerings, improve fraud detection mechanisms, and optimize risk management practices. To ensure that data-driven innovations benefit customers while protecting their rights and interests, fintech companies must handle privacy concerns, data security threats, and ethical issues related to the use of alternative data. A more accessible



and robust financial ecosystem is also fostered by industry standards and regulatory monitoring, which are vital in guaranteeing responsible data usage and boosting customer trust in fintech solutions.

### **3.2 Operationalization of Theoretical Constructs**

Several research techniques that were applied to enhance Bank Statement Analyser for SME loans are described in this chapter. This thesis chapter will go into great length on the data that were selected for the study, the difficulties encountered during data extraction, data transformation, data pre-processing, and the approaches utilized during model construction and assessment. Furthermore, it explains the rationale for the classification method selection in comparison to all other feasible approaches.

In order to support research aims and address research questions, a series of logical diagrams have been produced during this study. These diagrams offer a thorough explanation of the full sequence of procedures taken for obtaining the intended results. To help with a better understanding of the complete research technique and design, these diagrams are provided and explained in detail in the following sections. Subsequently, this chapter focuses on the large language model that was employed for the study and how it was highly beneficial in producing the research's results. Lastly, a number of difficulties were encountered when carrying out the research, and each of these difficulties was handled with appropriate justification and mitigating measures.

### **3.3 Research Purpose and Questions**

The objective of this study is to enhance a bank statement analyzer for SME lending proposed by the study (Mehta, 2020). This analyzer has the potential to generate insightful data that would assist sound credit decisions and lower lending risk to the SME community. A more robust and superior decision-making process would result in less risk. The quantity of transactions required ought to be sufficient to yield a proof of concept and to yield valuable insights that could be replicated for any number of transactions in the future. The

US Lending Market is the geographic area that this study takes into consideration. Large Language Model is used as an alternative for the classification of the banking transactions. Moreover, the model is also applied to India lending market to see if this would work for different geographical location for generalization.

This research study has the following mentioned objectives.

- To enhance the list of keywords used by the study (Mehta, 2020) for categorizing banking transactions of small and medium enterprises in the United States.
- To enhance the performance of the machine learning model, build in the study (Mehta, 2020) for categorizing banking transactions of small and medium enterprises in the United States.
- To perform the classification of banking transactions of SME into the above identified categories using novel AI that includes Large Language Models. Comparing the results of traditional AI (ML model) and LLM classification for the banking transactions.
- To apply the above study for different geographical location i.e. India lending Market to see if the same model has worked well for another location so that it could be generalized.

### **3.4 Research Design**

This section describes the research approach undertaken by the study (Mehta, 2020) and then would produce the enhanced design to improve the bank statement analyzer for United States. Furthermore, the same study could then be applied to India Lending market to see the applicability of the same for generalization.

#### **3.4.1 Elaboration to Approaches**

A previous study (Mehta, 2020) analyzed the bank statements of US small and medium-sized businesses. Small and medium-sized businesses have two options: either give the Fintech's access to their bank statements, or give permission to download them from Yodlee. This section described the two methods used to carry out the investigation. The

first method had to be used as there were no keywords or categories for transactions; instead, keywords had to be created, examined, and prepared for classification. This established a foundational vocabulary for keywords, and as time went on, this list would expand. Moreover, auto-learning or self-learning techniques that are not covered in this study might be used for the same keywords.

1,00,000 parsed transactions from 3,300 SMEs were examined in this study. To generate keywords for the purpose of classifying the transactions, all of the input data was utilized. Based on the specified keywords, these transactions were further categorized against additional classifications and groupings. In this study, the procedure of calculating metrics based on these categories is referred to as post classification.

Following the generation of keywords, they were mapped to classes and groups, leading to the creation of a master keyword data sheet. The input data is then classified using these keywords. Since rules have been embedded in the program, rule-based categorization has very high accuracy and would yield very accurate results for output variables. For these reasons, it was chosen. The rule-based engine's accurate classification of the data allowed it to be used for the machine learning technique.

The rule-based engine assigned keywords based on the following criteria. The master keywords list and the keywords found in a transaction's description are compared. It's possible that more than one keyword from the master keywords list matches the transaction. In this instance, the keyword's initial occurrence is examined. Classes and groups are assigned to this transaction from the master keywords file once the first keyword is detected and the transaction is mapped to that keyword.

The second method uses 1,00,000 labeled banking transactions as classified data as input data. Seventy percent of the classified data, which included labels, was used to train and construct the model. The multi-class classification model was constructed using Random

Forest, as was indicated in the modeling section. The model evaluation was then carried out using this model and 30% test data. This is one method that might be applied when keywords are established; the list of keywords may expand in the future as a result of lessons learned from the misclassifications.

Following the classification process using approaches I and II, post-classification techniques were used to calculate several metrics that are detailed later in this section. The final product of this research project is cash flow variables analysis.

The section describes above the approaches undertaken by the study (Mehta, 2020). The understanding of these approaches is mandatory for the current research and thus they have been elaborated above. The classification previously done using multi class classification mechanism via Random Forest machine learning method could be enhanced by using novel AI that is Large Language Model (Chat GPT 4) and specialized prompting engineering by employing prompt template and prompt designs.

### **3.4.2 Logical Flow Diagrams**

This section presented the logical flows of the two approaches used in previous study (Mehta, 2020). This is important to understand them so that enhancements could be proposed and executed as a part of this research.

#### **3.4.2.1 APPROACH I**

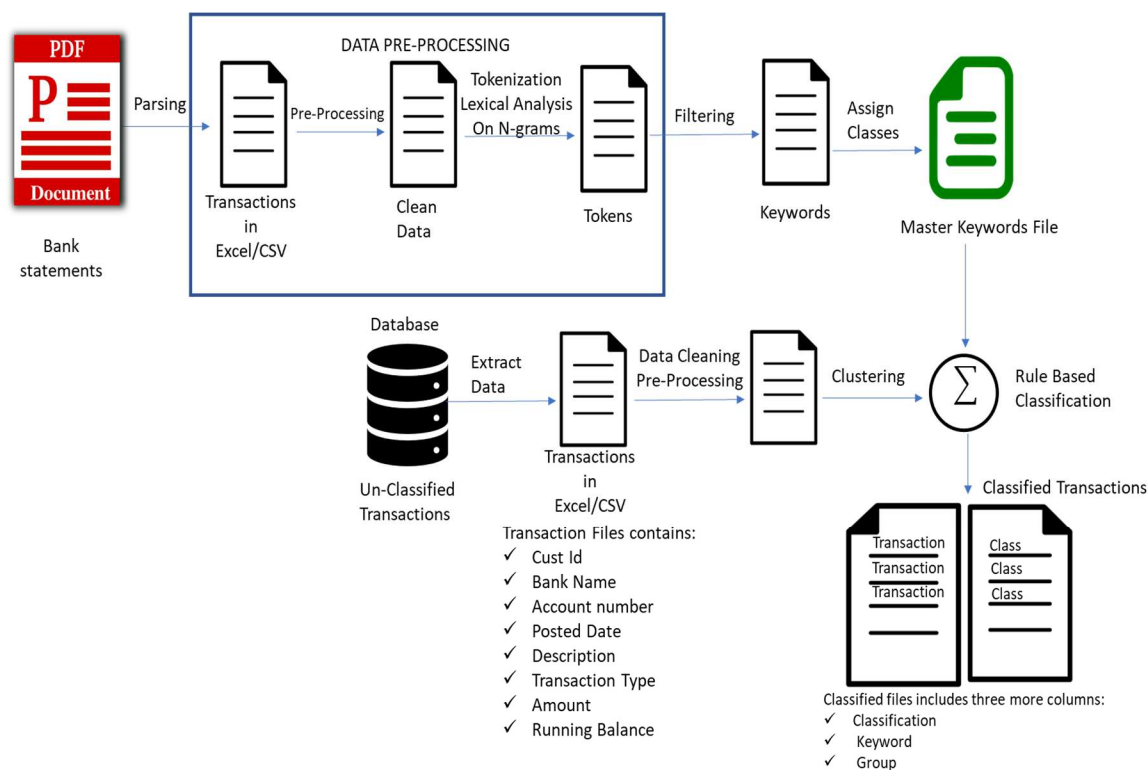


Figure 1 Process Flow Diagram

The diagram above illustrates how transaction-level data from small and medium-sized businesses' PDF bank statements is parsed and extracted into the CSV format for further analysis. This parser is outside the purview of this study because PDF bank statements include private information about the bank and its customers. Only transactions are visible after being extracted into Excel; all other data is deleted. Additionally, before being shared for this study, spreadsheets were anonymized. Data Selection in the context of research technique. For the purpose of this study, 1,00,000 anonymized, parsed transaction data were used.

EDA was carried out using univariate and bivariate analysis to develop the necessary comprehension of the data. Pre-processing of the data was done to eliminate special characters, numerals, and stop words. Additionally, for future treatments, transaction descriptions would be altered to lower case. For additional processing, the date column would be converted to Date Time format. Then, using the count vectorizer and n-grams

models, the transaction descriptions would be transformed into tokens. As a result, n-grams (where  $n = 1$  and  $2$ ) along with their frequency of occurrence in the transaction descriptions are combined to create keywords. In the master keyword sheet, the previously mentioned keywords are analyzed and mapped to classes and groups.

As described in the later portion of the chapter, processed transaction data from the previously mentioned phases are clustered using the DBSCAN clustering algorithm. The study's implementation section provides an explanation of the parameters selected. As stated in the previous phase, clustering is used to group similar transactions. Every cluster's frequently occurring transactions are chosen for classification. Every transaction in a cluster is classified using rules, and all of the transactions in that cluster are given the same keyword. The master keyword sheet consists of classes and groups that have been assigned.

For classified transactions, post classification is performed. It computes the important metrics and analyzes their trends. The following are the output variables that the post classification module created. The gross income derived from the business is known as revenue. Operating costs are the overall amount spent on running a business. Operating profit, also known as operating income, is the difference between total expenses and gross revenue. The ratio of operating income to revenue is known as operational profit margin. Debt Service Coverage Ratio is referred to as DSCR. It serves as an indicator for the availability of sufficient money to pay off debt. It is, thus, the operating income to debt service ratio. The acronym for debt-to-income ratio is DTI. This represents the simple ratio of debt to operating income. The total amount required to pay off debt, including principal and interest, is known as debt services. The entire debt services divided by the revenue is known as debt repayment per revenue. The Cash Flow Ratio calculates a customer's ability to pay off debt or liabilities with operational income.

### 3.4.2.2 APPROACH II

The study (Mehta, 2020) uses 1,00,000 transactions from the first method that have already been classified. When the banking transactions have already been classified, this method establishes the course of action that needs to be followed. Furthermore, this approach has certain limits concerning computation speed and performance, which will be further discussed later in this paper. The second method is illustrated in the following figure.

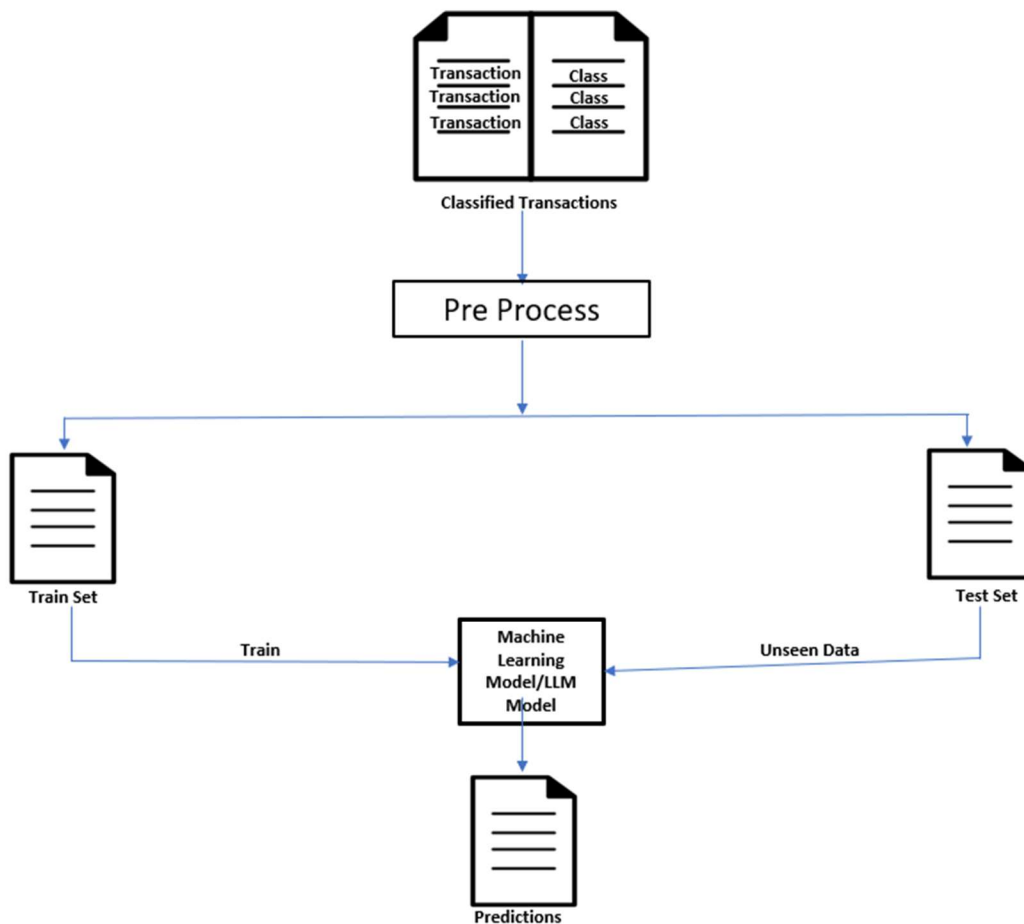


Figure 2 Machine Learning Model/LLM Model

As stated in method I, data pre-processing involves eliminating special characters, integers, and stop words from the transaction descriptions. Following that, transaction descriptions were changed or converted to numerical form. Train and Test data sets are separated. The

split ratio employed for the train and test datasets was 70 and 30, respectively, when random sampling was carried out. After training on the training set, the random forest model was saved. The implementation defines the parameters that are used. The previously mentioned model's performance was tracked and it was confirmed using a test set. Accuracy was utilized to evaluate performance, as was indicated in this chapter's section on model evaluation. To determine accuracy, the produced categories were thus compared to the real labels. The test dataset was subjected to the post-classification procedure in order to produce important metrics and trends. In this study, the results of the two previously described methodologies are explained in detail along with the enhancements proposed as a part of this study. In the current study, instead of just ML model, LLM model with intensive Prompting techniques is used for the classification of the bank statement transactions.

### **3.4.3 Tools and Technologies**

The techniques and technologies employed in this research are presented in this section. In this study, the following technologies and tools were employed.

#### **3.4.3.1 PYTHON**

Python, a popular programming language for data science, and its many libraries will be used in this study. In this study, Python version 3.7 is employed. The following are specific Python libraries and tools that will be used in this project. With Pandas, one can do end-to-end large-scale data analysis and data product development since it is a robust open-source library that offers the necessary data structures and data analysis skills. Numerous mathematical functions and multi-dimensional arrays are supported by the NumPy library. Modules for regular expressions are often included with the "Re" library for string manipulation and search. The Matplotlib package is used for visual aids. A Python machine learning library called Sklearn, sometimes known as SciKit Learn, offers a number of supervised and unsupervised learning techniques. It also offers Count Vectorization and TF-IDF techniques. NLTK (Natural Language Toolkit) is used in this research for dealing



with Text Mining or for providing methods for Text processing. Numerous techniques for computing statistics are available in the Python statistics library.

#### 3.4.3.2 JUPYTER NOTEBOOK

Code snippets are created and results are visualized using a Jupyter notebook.

#### 3.4.3.3 MICROSOFT EXCEL

The relevant classes and groupings were compiled with the aid of Microsoft Excel. The execution of code and analysis on a 64-bit Microsoft Windows desktop computer with 32 GB RAM was necessary because the processing of 1,00,000 transactions data was not smooth and required a lot of computing time and memory. GPUs could also be used in other cloud environments, such as Google Colab, a free cloud platform for machine learning research.

#### 3.4.3.4 CHAT GPT

OpenAI's ChatGPT is an advanced conversational AI model built on the Generative Pre-trained Transformer (GPT) architecture. The most recent version, GPT-4, expands on the capabilities of its predecessors by using deep learning methods. ChatGPT makes use of the Transformer neural network design, which requires sequential data, such as text. The algorithm learns to predict the next word in a sentence by being pre-trained on a sizable corpus of internet material. It gains a wide grasp of language during this pre-training phase, including knowledge of syntax, world facts, and even some cognitive abilities.

The model transforms the text into numerical representations, which it then processes using the Transformer architecture's several layers. With each layer, the comprehension of the text is improved by taking into account the context that the surrounding words provide. These enhanced representations are decoded back into text that is understandable by humans.

Despite ChatGPT's strength, there are certain drawbacks. It occasionally yields answers that seem convincing but are inaccurate i.e. a condition known as "hallucination." Explainability of the outcome is also a challenging task. Responsible and Ethical AI is the important consideration as far as Chat GPT is concerned.

### **3.5 Population and Sample**

This analysis took into account the parsed banking transactions of small and medium-sized US and India Lending businesses. The dataset's key details that contributed to the necessary data interpretation are listed below. There were 1,00,000 transactions in the dataset, and 3,300 small and medium-sized businesses were taken into consideration for the US market. There are 20,000 transactions, and 450 SMEs for India market. Date period taken into consideration: December 2013–October 2019 (United States) and April 2021–March 2023 (India). Cust Id, Account number, Post-date, Description, Transaction Type, Amount, and Running Balance are the fields in the dataset used for the calculations of features of model.

### **3.6 Participant Selection**

This section explains the process of preparing and dividing the data in order to conduct this study. First, web crawlers and scrapers are used to extract the data from the PDF file into Excel in the correct format. Similar fields or data can be found on every bank statement, such as account information, which includes account details including the account number, customer ID, address, branch, and name of the account holder. The information on transaction categories, transaction descriptions, amounts, running balances, and value dates would also be included in the transaction statement. Later this section also contains the various prompts used in a prompt template for the generation of the responses from the Large Language Model.

#### **3.6.1 Data Preparation**

To prepare the data for modeling, data preprocessing was necessary. For this study, the data preparation for each of the major procedures is explained below. This study is divided

into four modules: keywords generation, clustering and rule-based classification, machine learning and LLM classification and post classification. These modules' specifics are discussed later in this chapter. Let's outline the data preparation tasks derived from these modules.

### **3.6.2 Data Preparation for Keywords Generation Module**

The descriptions of transactions were changed to lowercase. The transaction descriptions were cleaned up by eliminating many spaces. Special characters and numbers were eliminated from the transaction descriptions. The transaction descriptions were edited to exclude stop words in English. Tokenization of transaction descriptions was done using N-Grams (n=1, 2 & 3). Following this, word tokens consisting of single, double, and triple word combinations were produced called as keywords.

### **3.6.3 Data Preparation for Clustering and Rule-Based Classification Module**

The transaction description includes several spaces that are removed. The transaction description was cleaned so that it contains only letters. The transaction descriptions were converted to lower case. The stop words are removed from the transaction descriptions. Tokenization of transaction descriptions was done using N-Grams (n=1,2, & 3). TF-IDF vectorization was used to convert text (the tokens above) into numeric vectors. Transactions were prepared for clustering using the previously created numeric vectors after completing all of the previous stages.

### **3.6.4 Data Preparation for Machine Learning Classification**

The transaction description was cleaned up by removing isolated characters, numerals, special characters, and multiple x's. It was decided to convert the transactions description column to lowercase. Stop words in English were eliminated from transaction descriptions. Using the TF-IDF technique, the processed description was converted from text to numeric. For each transaction type, dummies were made. A single data frame had concatenated

dummy variables for Transaction Types, TF-IDF vectors, running balances, and transaction amounts.

### **3.6.5 Data Preparation for Large Language Model Classification**

Preparing prompts for multi-class classification using a Large Language Model (LLM) like GPT, Gemini etc. involves designing queries that effectively helps in performing the desired classification from the model. The key is to structure the prompts in a way using various prompt designs and prompt templates that clearly presents the categories and ensures the model understands the task. Below is an explanation of how to design effective prompts using various prompt templates and designs for multi-class classification using the Large Language Model using Chat GPT. However, in this study I have used the various but not all the possible prompt designs due to the time constraint and the bigger objective was not to test prompts but verify the possibility of effective prompting for the use of Chat GPT for the classification task.

#### **3.6.5.1 DIRECT PROMPTS**

Simple direct inquiry prompts ask the model to categorize an input into one of the pre-established categories. When the task is clear and the categories are distinct, this technique performs effectively.

Prompt Template:

Categorize the following banking transaction based on the transaction description into one of the following categories: “Revenue”, “Expenditure”, “Liability”, “Debt Service”, or “Transfer”.

Transaction Description: [Input Text]

### 3.6.5.2 MULTIPLE-CHOICE PROMPT

Template:

Read the transaction description: [Input Text]. Which of the following categories does this transaction belong to?

- A) Revenue
- B) Expenditure
- C) Liability
- D) Debt Service
- E) Transfer

### 3.6.5.3 CONTEXTUAL CLASSIFICATION PROMPT

Template:

Based on the transaction description provided, categorize it as one of the following:

“Revenue,” “Expenditure,” “Liability,” “Debt Service,” or “Transfer.”

Description: [Input Text]

Context: [Reference Text]

### 3.6.5.4 INSTRUCTIONAL PROMPT

Template:

You are an expert in financial analysis or cash flow analysis. Given the following transaction description, classify it into one of the categories: “Revenue,” “Expenditure,” “Liability,” “Debt Service,” or “Transfer.”

Transaction: [Input Text]

### 3.6.5.5 EXAMPLE-BASED PROMPT

Template:

Here are some examples of categorized transactions:

1. "Salary payment received" -> Revenue
2. "Loan repayment made" -> Debt Service
3. "Purchase of office supplies" -> Expenditure
4. "Funds transferred to savings account" -> Transfer
5. "Interest on loan" -> Liability

Now, categorize the following transaction: [Input Text]

Or the above examples could be given as Few Shot Examples in a template

### 3.6.5.6 CHAIN-OF-THOUGHT PROMPT

Template:

Let's break down the following transaction step by step: [Input Text]. First, identify the key elements, then determine the most appropriate category from "Revenue," "Expenditure," "Liability," "Debt Service," or "Transfer."

### 3.6.5.7 BINARY DECISION TREE PROMPT

Template:

First, ask if the transaction increases assets or reduces liabilities (Revenue or Transfer). If it reduces assets or increases liabilities, it could be Expenditure, Liability, or Debt Service. Categorize the following transaction description:

Transaction: [Input Text]

#### 3.6.5.8 ROLE-BASED PROMPT

Template:

As a Credit Risk Analyst, you need to categorize the following transaction. Choose the correct category from “Revenue,” “Expenditure,” “Liability,” “Debt Service,” or “Transfer.”

Transaction Description: [Input Text]

I could use above BOT Persona or User Persona.

#### 3.6.5.9 SCENARIO-BASED PROMPT

Template:

Imagine you are analyzing a bank statement, and you come across the following transaction: [Input Text]. Which category best describes this transaction: “Revenue,” “Expenditure,” “Liability,” “Debt Service,” or “Transfer”?

#### 3.6.5.10 ELIMINATION-BASED PROMPT

Template:

Consider the following transaction description: [Input Text]. Eliminate the categories that don't fit until you find the most appropriate one from “Revenue,” “Expenditure,” “Liability,” “Debt Service,” or “Transfer.”

#### 3.6.5.11 CONFIRMATORY PROMPT

Template:

The transaction described as [Input Text] seems to be [Proposed Category]. Is this correct? If not, select the most appropriate category from “Revenue,” “Expenditure,” “Liability,” “Debt Service,” or “Transfer.”

#### 3.6.5.12 COMPARISON-BASED PROMPT

Template:

Compare the following transaction description with the categories: “Revenue,” “Expenditure,” “Liability,” “Debt Service,” or “Transfer.” Which category does it align with best?

Transaction Description: [Input Text]

Not all of above but most of them have been clubbed into the prompt design to create the most appropriate prompt template for my study. The actual templates used and the results are described in the later sections.

### **3.6.6 Data Preparation for Post Classification Module**

For the revenue, cash flow, and DTI outcome variables, data ranges or distribution ranges were established. To extract the month-year values from the post date, the data type of the post date was converted to Date-Time format. The module essentially computes various intermediate values and final cash flow variables after completing those two stages.

### **3.7 Data Partitioning**

The train and test sets were created by randomly dividing the processed data. This phase was only carried out for Approach II, where the Random Forest Model was used to model



the data. 70% of the processed data that was randomly chosen to serve as training data was used to train the model. Thirty percent of the processed data that was randomly chosen referred to as test data was used to test or validate the model. Using a rule-based classifier, I assigned keywords to each transaction based on the entire 100% of input data that I used for keyword generation.

### **3.8 Instrumentation**

This section explains the several modeling techniques utilized in the study and provides justification for choosing these particular techniques over others.

#### **3.8.1 Transactions Clustering**

Similar transactions were grouped together using the clustering technique. The same category or keyword was then allocated to the transaction groups. The input for the clustering is the combined transactions for each kind (credit/debit) for every client. The transaction descriptions are then vectorized using TF-IDF in conjunction with n-grams (n ranges from 4 to 10). Transactions using debit and credit had different parameters. After vectorization, clustering was performed using the DBSCAN algorithm as described in (DBSCAN - Wikipedia, 2020). A smaller selection of these representative transactions was utilized for classification, and one representative transaction was selected from each cluster.

DBSCAN stands for Density-based spatial clustering of applications with noise, as stated in (DBSCAN - Wikipedia, 2020). Based on the distance between the points and the minimum number of points in a dense zone, this unsupervised technique groups the points. Refer to the diagram below to understand how DBSCAN operates. The data points in the diagram below are represented by red dots, outliers by yellow dots, and noise by blue dots. The distance between any randomly selected red data point and the closest points is calculated. They will form a single cluster if the distance is within the predetermined

minimum distance (eps). The procedure is then repeated, taking into account the subsequent red data point.

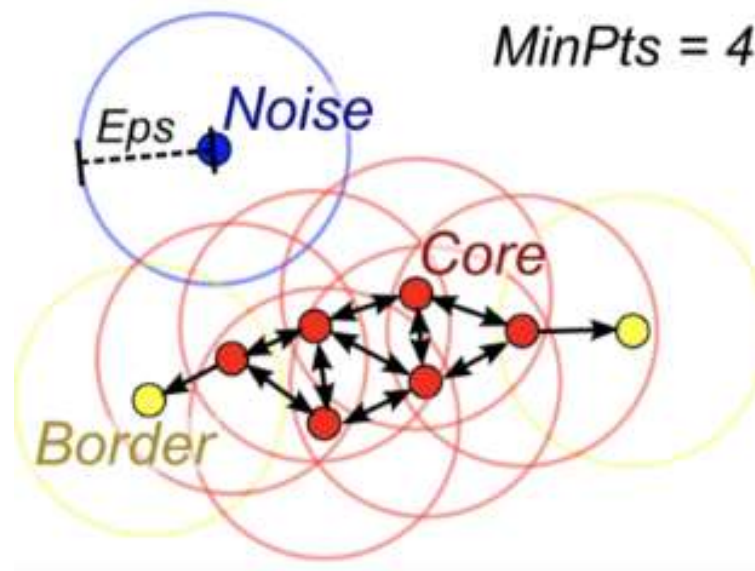


Figure 3 DBSCAN clustering process. (DBSCAN - Wikipedia, 2020)

The acronym for Minimum Points required in a cluster is MinPts. This value is 4 in the diagram above. The number of dimensions in the dataset determines the optimal value of this statistic.  $MinPts = D - 1$  if  $D$  is the number of dimensions in the dataset. This parameter has a minimum value of 3.

Eps represents the distance between the data points, or epsilon. One could obtain this value by utilizing the so-called k-distance graph. The graph displays the distance in descending order of distance values plotted against  $MinPts - 1$  nearest neighbors. The graph's elbow appears at the ideal value of epsilon. This is the value selected. The only value for  $K$  in the k-distance graph is  $K = MinPts - 1$ . The minimum distance is in eps. A high number of clusters would arise from a very small value of Eps, and clusters would merge as a result of a big value, resulting in the majority of the points lying in the same cluster. Euclidean distance is the distance function employed in this study.

Additional important techniques for clustering include K-means Unsupervised Learning. The present study did not employ K-means clustering for the following reasons. Unlike K means, DBSCAN does not require the number of clusters to be specified. By employing DBSCAN unsupervised learning, clusters of any shape might be discovered. Unlike K means, the DBSCAN approach is unaffected by outliers. The DBSCAN application requires the fewest possible parameters.

### **3.8.2 Transactions Classification using Machine Learning**

Machine learning encompasses a range of classification techniques, such as SVM, Random Forest, Decision Trees, Logistic Regression, and Extreme Gradient Boosting. The following facts served as the foundation for the classification selection model. The complexity of the dataset rises following the text to numeric conversion, which places restrictions on the application of logistic regression. Due to the sequential structure of the procedure, boosting will take longer to execute for large dimensionality and high volume datasets. The same is true for SVM; it would not be the optimal option for this research due to its high processing resource requirements. Compared to Random Forests, Decision Trees may produce biased and overfitting results. Random Forest is the machine learning method that was employed in this investigation. By using bagging to aggregate the decision tree results at the end, random forests generate objective findings. In comparison to the other techniques discussed, Random Forest runs decision tree models in parallel, which reduces computing complexity and improves performance through quick training. Because it splits at a node using a subset of features, it works well with high dimensionality data. How does this operate and what is Random Forest? From the (Yiu, 2019), this was understood. An ensemble method called Random Forest makes use of bagged decision tree models. A bootstrap aggregation is called bagging. By "bootstrap," we imply that data selection for each decision tree model would take into account a random sampling of data with replacement. The final result is produced by adding the outputs from all of the models. At each node, a random selection of features is used for splitting, and homogeneity measurements like entropy or the Gini Index are used to determine spitting.

### **3.8.3 Transactions Classification using Large Language Model**

A powerful artificial intelligence system known as a Large Language Model (LLM) trains on vast volumes of text data to understand and produce human language using deep learning techniques. These models, like GPT-4, are especially useful for a variety of language-related tasks, such as the classification of financial transactions, because they are made to capture complex linguistic patterns, context, and semantics. Within this framework, LLMs are able to examine transaction descriptions and group them into pre-established types such as Transfer, Expense, Revenue, Liability, and Debt Service. For this objective, LLMs have a number of advantages over conventional Machine Learning (ML) models. Because LLMs are particularly good at reading context and nuances in text, they don't require as much feature engineering as ML models do. Additionally, they are more adaptable, able to handle a variety of tasks with little modification, and able to take advantage of transfer learning, which enables them to adjust to particular jobs with relatively less new data. However, there are significant disadvantages to classification with LLMs. These models are expensive and at times not feasible for real-time applications since they demand a lot of processing power. Additionally, LLMs operate as "black boxes," which means that it is difficult to understand how they make decisions. This presents problems in industries like finance where transparency is essential. Furthermore, when fine-tuned on smaller datasets, LLMs may overfit and sustain biases identified in the training data, raising questions regarding accuracy and fairness. It all depends on a thorough examination and evaluation of the potential and constraints of LLMs in NLP and Novel AI use case like this study.

### **3.8.4 Evaluation**

This was a challenging section of the study because it is merely a proof of concept and there hasn't been much research done on the classification of bank transactions in the literature. There was no labeled data available. In this study, labels were created and manually checked. In this study, labels refer to the terms used to categorize the transactions.

This was required in order to generate different metrics. For instance, revenue depends on cash flow, which is dependent on a number of keywords or certain transaction categories, such as "Deposit," "Sales," and so on. The validation and assessment of this study was carried out using the following procedures.

#### 3.8.4.1 KEYWORDS VALIDATION

To organize transactions into important categories, keywords were developed. Based on the previously described procedures, keywords were generated from data transactions via Natural Language Processing techniques such as count vectorizer, N-Grams, and TF-IDF scores. Following that, these keywords were manually examined using the Banking domain expertise relevant to SME financing. These keywords were compiled into a master sheet along with a further categorization and rating. After that, this document was consulted and utilized in the transaction classification process. A section later explains this complete sequence.

#### 3.8.4.2 TRANSACTION CLUSTERS VALIDATION

The manual validation of transactions that were grouped using the DBSCAN process involves examining the descriptions of some of the transactions that belong to the same cluster.

#### 3.8.4.3 CLASSIFICATION VALIDATION FOR ML AND LLM MODELS

This study used two approaches, which are discussed later in this chapter. Rule-based classification was employed in the first method, and machine learning/ Large Language Model classification was used in the second. Using a classification algorithm, all of the transaction data was finally matched to the keywords. The output used in this study to carry out and record the results of the second technique was referred to as labelled classified data. Given that every keyword was examined and analyzed, machine learning might use this dataset in the future to categorize new SME lending transactions.

The first keyword match between the keywords in the transaction description and the corresponding keywords in the master keyword file was the basis for the rule-based classification that was employed. Because rules were incorporated into the program to categorize the transactions, the findings have been extremely accurate. In the second approach, the rule-based technique and the classification results from the machine learning method were compared in order to assess the effectiveness of the machine learning method and to determine areas for future modeling improvements.

Accuracy was chosen for the machine learning classification method's performance evaluation for several reasons. After classification, the keywords were assigned to the validation set of data. The actual keywords in the categorized input dataset were compared with these expected keywords or classes. This comparison led to the creation of a confusion matrix, and since we were equally interested in each class, accuracy served as the best metric by indicating the percentage of keywords that were correctly allocated to the validation set of transactions.

### **3.8.5 Model Implementation**

The entire dataset was cleaned up and EDA was performed. Data preparation was defined separately for each important procedure used in this investigation. Similarly, modelling or model implementation differed for four critical processes, as detailed below. This research was separated into four modules: Keyword generation, Clustering and rule-based classification, Machine learning and Large Language Model classification and post-classification. The model implementation in each of the four modules listed above is discussed later in this paper.

## **3.9 Data Collection Procedures**

In order to enhance bank statements analyser for SME lending, PDF bank statements data used in the previous study (Mehta, 2020) is needed along with the transaction level data for India Lending Market for stress testing. The bank statement data utilized in this study

is described as follows, whether scanned or native PDFs, were to be analyzed to produce transaction-level data. Transaction level data must include transaction descriptions, kinds, account information, and amounts. The transaction date was required since it served as the foundation for calculating numerous metrics. The two main transaction kinds are credit and debit. It describes the nature of the financial transaction. Bank statements are very sensitive since they contain account-level information, customer personal data, and bank information. There were various obstacles in obtaining this secure data and anonymizing it for research (Mehta, 2020). According to this study, anonymizing the bank statement pdfs didn't function since after anonymization, the parser built to extract information from bank statements stopped functioning because the PDF attributes were altered by the anonymization procedure.

In the study (Mehta, 2020), PDF bank statements were obtained from 3,300 small and medium-sized firms in the United States. The same data is reused in current study along with an additional similar data for India lending Market. The parser was designed to parse these bank statements and retrieve transaction-level data as well as account-level information in Excel format. It was simple to anonymize the excel file by inserting random values into important columns. The following columns are anonymized in the parsed data by assigning random values. For customer identification, a unique 6-digit random combination is employed to ensure anonymity. Except for the last four digits of the account number, all other digits are substituted with 'x' to ensure anonymity. The transaction description was left unchanged because it served as the foundation for the establishment of keywords for transaction categorization.

In addition to this, for using large language models for the multi class classification several prompt templates are prepared defined in sections later as a part of data creation process. Transaction descriptions are used in prompt templates for classifying the transactions into the cash flow category.

### **3.10 Data Analysis**

This section describes the various components of data analysis phase of the study (Mehta, 2020) and the same data analysis is also performed for current study for enhancing the bank statement analyser along with the stress testing with India Lending Market.

#### **3.10.1 Data Pre-processing**

Data pre-processing includes duplication removal, data type correction, missing value treatment, outlier treatment, normalization or standardization, categorical data encoding, and feature generation (feature engineering).

The data used for this study was parsed from pdf bank statements, therefore it was rather clean and free of duplicates and missing numbers. Outliers in numeric categories such as quantity and running balance were not treated for this study since they would change the original data and give conflicting results. For each customer, metrics such as revenue, operational cost-profit, and others were generated, and they were heavily dependent on the amount fields. During data pre-processing, the Post Date column was converted to Date Time format so that it could be broken down to month and day levels for analysis.

This study pays more attention to the pre-processing of the text data that was included in the dataset. The transaction description is a crucial component of the data that is used to generate keywords and classify transactions. Text pre-processing was accomplished using lexical analysis. While there are several approaches to text pre-processing in lexical analysis, we have employed some of them, as described below. First, transaction descriptions are changed to lowercase characters. if Transaction Statement = "MISCELLANEOUS FEE REVERSAL," then the result of converting to lowercase is "miscellaneous fee reversal."

The most often used words in English are stop words, such as "is," "and," "this," etc. The English stop words were eliminated from the transaction description column. It was also



done to eliminate special characters and numerical data from the transaction description. NLP uses stemming and legitimization techniques to reduce terms such as "racing," "racer," and so on to their root word, "race." However, as the information comes from bank statements and includes a variety of financial terms along with acronyms and terminology related to finance, these procedures weren't applied.

Tokenization was used to divide the text into individual words or tokens. The transaction description was divided into a list of words using white spaces. Tokenization was carried out since it helped in text analysis. For example -

Transaction Statement: "reversal of miscellaneous fees"

result of tokenization = ["miscellaneous", "fee", "reversal"].

This entire process as described above is repeated for current study as well for generating additional key words and using it for the enhancement of the bank statement analyser generated by the study (Mehta, 2020).

### **3.10.2 Data Transformation**

There are numerous data transformation methods used in study (Mehta, 2020) and are also used in the current study for enhancing the result of the study (Mehta, 2020) along with the application of the large language models. This section only highlights those transformations that have been used in this study for reproducing a bank statement analyser for enhancement. This study employed a number of techniques to convert the text in transaction descriptions into numerical form. The details for these techniques are listed below. The machine can comprehend numbers in the end, and different techniques were used to translate word tokens into numeric form for modeling purposes. These techniques essentially assign words probability or frequencies. Here are a few of the ones that were used in this study.

### 3.10.2.1 COUNT-VECTORIZER

Tokenization is the process of extracting words from text, as stated in the study (Heilenreich, 2020). For usage in the machine learning technique, these words or tokens must be encoded as numbers; this process is known as vectorization or feature extraction. According to a study (Egesa, 2016), count-vectorizers extract words from text documents and create a vocabulary by storing words along with their frequency of occurrence in text documents. This process is known as tokenization. This vocabulary might then be used by count-vectorizer to encode new documents. The length of the encoded or output vector matches this vocabulary, and it contains a count that indicates how many times each word has appeared in a new text. Punctuation is removed and words are converted to lower case.

### 3.10.2.2 TF-IDF SCORING

According to the study, the TF-IDF score is calculated as the product of "TF," or word frequency, and "IDF," or inverse document frequency (Maklin, 2019). This is a significant metric that essentially determines a word's relative importance within the corpus. As a result, it checks the total number of documents that contain the word in addition to the number of times it appears in a document. It offers distinct terms in this manner, and those words would have high TF-IDF ratings. The number of times a term appears in a document divided by the total number of terms inside the document is called as term frequency.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Figure 4 Term Frequency as defined in (Maklin, 2019)

The number of documents where a term would have appeared is known as document frequency (df) for that word. The ratio of the total number of documents (N) to the

document frequency is known as the inverse document frequency, or idf. This would be high for terms that are uncommon.

$$idf(w) = \log\left(\frac{N}{df_t}\right)$$

Figure 5 Inverse Document Frequency as described in (Maklin, 2019)

"idf" stands for inverse document frequency in the formula above. "df" stands for document frequency, while "N" is the total number of documents. The term or word in question is "t." Consequently,

$$\text{TF-IDF score} = \text{tf} * \text{idf}$$

### 3.10.2.3 N-GRAMS METHOD

The N-grams method was the next technique applied to text pre-processing. The n-gram approach outlined in Srinidhi (2019) would yield n-word sequences. The bi-gram approach, for our transaction description "monthly service fee deducts," for instance, would provide the following word sequences: "monthly service," "service fee," and "fee deducts." However, the 3-gram model would have produced the following results: ["monthly service fee", "service fee deducts"]. N-gram models essentially calculate the likelihood of the final word based on the preceding word's sequence. Because keywords for transaction classification might have a maximum of three words, such as "ATM DEPOSIT FEE," which sets it apart from all other types of deposits transactions and used in our research.

### 3.10.3 Data Cleaning

This part of the document includes the checks to verify the veracity of the data. Because the data used in this study came from bank statements with a high degree of authenticity, it was clean data.

#### 3.10.3.1 DUPLICATES CHECK

The dataset had no duplicates. Every row that represents a transaction is distinct.

#### 3.10.3.2 MISSING VALUES

No missing data value was found for any of the fields present in the input dataset.

#### 3.10.3.3 OUTLIERS

Outliers found for the amount and running balance fields were not treated as data was genuine and these fields were critical for generating output cash flow variables. If outliers in these numeric fields were treated, this would have generated inconsistent cash flow variables.

### **3.10.4 Fixing Data Types**

Date column was converted to Date Time format.

### **3.10.5 Exploratory Data Analysis**

The primary goal of this study was to analyze and derive cash flow variables after classification; hence, exploratory data analysis was used to develop a foundational understanding of the entire set of data. Data analysis was done, including bivariate and univariate data analysis. Developing a statistical understanding of a single variable, such as its variance, distribution, etc., is essentially the goal of univariate analysis. Understanding the relationship between two variables is performed via Bivariate analysis.

Merchant ID	# Bank Account		Merchant ID	# Bank Account
321380	8	↑	308148	1
318465	7		313082	1
316728	6		314036	1
312745	5		321218	1
321356	5		321671	1
326062	3		325142	1
310443	2		325330	1
314559	2		326050	1
321146	2		328212	1
323253	2		329803	1
			330698	1

Figure 6 Number of Bank Accounts of Merchants

Merchant ID's 316728 & 321380 have bank accounts in 2 different banks and have higher number of bank accounts, 6 and 8 respectively. Merchant ID 318465 has 7 bank accounts within a single bank. ~70% of the merchants have either 1 or 2 bank accounts.

*Bank Accounts with Few Months of Data*

ID	Bank Account ID	Number of Months
312745	12837	1
318465	13421	2
321380	14630	2
321380	14631	2
321380	14632	2
321380	14633	2
326062	14045	2
318465	13420	3



*Bank Accounts with Higher Number of Months of Data*

ID	Bank Account ID	Number of Months
308148	12460	10
313082	13233	10
314036	13226	11
325330	15002	11
321218	14779	13
321671	14629	13
328212	14981	13
329803	14339	13

Figure 7 Number of Months of each Bank Account

Many Merchants with a higher number of bank accounts, have some bank accounts which are not operated often and their data is available only for less than 3 months with really low number of transactions occurring in those accounts

Majority of Merchants who submitted/have only one bank account is available for higher number of months and have decent number of transactions occurring in those accounts as well.

ID	Bank Account ID	# Months	# Credit Transactions	# Debit Transactions	Total Credit Amount	Total Debit Amount	Monthly Avg Credit Transactions	Monthly Avg Debit Transactions	Monthly Avg Credit Amount	Monthly Avg Debit Amount	# Bank Accounts of Merchant
312745	12837	1	1	0	25	0	1	0	25	0	5
318465	13421	2	0	2	0	5	0	1	0	3	7
326062	14045	2	0	2	0	51	0	1	0	25	3
318465	13420	3	4	4	50	55	1	1	17	18	7
312745	12835	5	7	11	16054	16054	1	2	3211	3211	5
312745	12836	5	3	5	19056	19035	1	1	3811	3807	5
314559	13273	7	8	4	5700	6080	1	1	814	869	2
316728	13235	11	17	3	3175	3175	2	0	289	289	6
321356	15147	11	21	12	2631	2652	2	1	239	241	5
321356	15148	11	19	10	2200	2246	2	1	200	204	5
323253	13877	11	17	7	20358	20112	2	1	1851	1828	2
310443	12655	12	8	23	16400	17767	1	2	1367	1481	2

Figure 8 Analysis of Credits, Debits & Monthly Averages of Bank Accounts

There are ~12 accounts which are dormant and have hardly 2 or 3 transactions taking place every month. All of these accounts are of Merchants who have submitted multiple bank statements and they do not transact frequently through these accounts like Merchant 312745 has 5 bank accounts among which 3 are dormant and really few transactions are taking place through these accounts.

ID	Bank Account ID	# Months	# Credit Transactions	# Debit Transactions	Total Credit Amount	Total Debit Amount	Monthly Avg Credit Transactions	Monthly Avg Debit Transactions	Monthly Avg Credit Amount	Monthly Avg Debit Amount	# Bank Accounts of Merchant
323253	13879	8	291	418	956059	970343	36	52	119507	121293	2
326050	14206	8	190	459	288066	271316	24	57	36008	33914	1
326062	14046	8	431	837	459574	456551	54	105	57447	57069	3
314036	13226	11	276	1985	445920	432975	25	180	40538	39361	1
325330	15002	11	312	257	82662	80230	28	23	7515	7294	1
326062	14044	11	421	604	382063	390477	38	55	34733	35498	3
321356	15146	13	389	395	1700025	1737552	30	30	130771	133658	5
321380	14637	13	2047	348	1234885	1239360	157	27	94991	95335	8
321671	14629	13	642	1788	312137	307190	49	138	24011	23630	1
328212	14981	13	282	345	177403	174828	22	27	13646	13448	1
330698	14374	13	346	906	10940565	10929049	27	70	841582	840696	1

Figure 9 Analysis of Credits, Debits & Monthly Averages of Bank Accounts Part 2

There are 11 accounts which are quite active and have high number of both debit and credit transactions taking place every month. Merchant 326062 submitted three accounts, 2 of those accounts are quite active with bank account 14046 being the most used account among the two. The remaining account of the merchant is not used much and rarely transactions happen on that account. Merchant 330698 submitted only one account but its heavily transacted and monthly transactions amounts are almost 20 times the average transactions amounts for the rest of the merchants.



ID	# Months	# Credit Transactions	# Debit Transactions	Total Credit Amount	Total Debit Amount	Monthly Avg Credit Transactions	Monthly Avg Debit Transactions	Monthly Avg Credit Amount	Monthly Avg Debit Amount	# Bank Accounts of Merchant	Total Credits minus Total Debits
321380	13	4939	773	2722094	2676941	380	59	209392	205919	8	45153
318465	9	369	1694	231849	229561	41	188	25761	25507	7	2288
316728	13	298	2314	2463020	2415844	23	178	189463	185834	6	47176
312745	13	160	600	719438	713251	12	46	55341	54865	5	6187
321356	13	697	788	4100524	4142485	54	61	315425	318653	5	-41961
326062	11	852	1443	841637	847078	77	131	76512	77007	3	-5441
310443	12	35	246	325133	346662	3	21	27094	28889	2	-21529
314559	13	143	284	527618	519814	11	22	40586	39986	2	7804
321146	9	121	454	105932	100592	13	50	11770	11177	2	5340
323253	12	308	425	976417	990455	26	35	81368	82538	2	-14038

Figure 10 Merchant Level Analysis (1/2)

Merchants submitting more than 4 bank accounts have a turnover (total credits) which is ~1.5 times of the rest of the merchants. For such merchants we should carefully analyze the transactions which are taking place as there may be transfers within the bank's accounts of one merchant. Merchant 321380 and 316728 have quite healthy profits in the bank account as total credits exceed total debits by 45k and 47k respectively. Merchant 321536 is not in good condition as we see total credits are lower than total debits by 42K.

ID	# Months	# Credit Transactions	# Debit Transactions	Total Credit Amount	Total Debit Amount	Monthly Avg Credit Transactions	Monthly Avg Debit Transactions	Monthly Avg Credit Amount	Monthly Avg Debit Amount	# Bank Accounts of Merchant	Total Credits minus Total Debits
308148	10	172	320	66611	61800	17	32	6661	6180	1	4811
313082	10	112	1026	547025	558877	11	103	54703	55888	1	-11852
314036	11	276	1985	445920	432975	25	180	40538	39361	1	12945
321218	13	57	849	1444858	1466503	4	65	111143	112808	1	-21645
321671	13	642	1788	312137	307190	49	138	24011	23630	1	4946
325142	8	115	876	472618	463603	14	110	59077	57950	1	9014
325330	11	312	257	82662	80230	28	23	7515	7294	1	2432
326050	8	190	459	288066	271316	24	57	36008	33914	1	16750
328212	13	282	345	177403	174828	22	27	13646	13448	1	2575
329803	13	116	655	219079	212388	9	50	16852	16338	1	6692
330698	13	346	906	10940565	10929049	27	70	841582	840696	1	11516

Figure 11 Merchant Level Analysis (2/2)

Merchant 308148 and 325330 have really low annual turnover which is ~ 80K to 90K in their bank accounts leading to lower overall total profits. Merchant 330698 has an annual turnover of \$1.1 Mn which is almost 13 times of the other merchants present in the data.

Industry	# Months	# Credit Transactions	# Debit Transactions	Total Credit Amount	Total Debit Amount	Monthly Avg Credit Transactions	Monthly Avg Debit Transactions	Monthly Avg Credit Amount	Monthly Avg Debit Amount	# of Merchants	Total Credits per merchant
Accommodation and Food Services	13	1214	2222	1196314	1180193	93	171	92024	90784	3	398771
Agriculture, Forestry, Fishing and Hunting	13	57	849	1444859	1466503	4	65	111143	112808	1	1444859
Construction	15	449	3215	3007232	2974893	30	214	200482	198326	3	1002411
Educational Services	13	282	345	177403	174828	22	27	13646	13448	1	177403
Finance and Insurance	13	4939	773	2722094	2676941	380	59	209392	205919	1	2722094
Health Care and Social Assistance	13	715	2600	11172414	11158610	55	200	859416	858355	2	5586207
Information Technology	13	143	284	527618	519814	11	22	40586	39986	1	527618
Other Services (except Public Administration)	13	812	1664	4573142	4606088	62	128	351780	354314	2	2286571
Professional, Scientific, and Technical Services	15	272	1626	1266463	1272128	18	108	84431	84809	2	633232
Retail Trade	14	1659	4909	1923069	1911443	119	351	137362	136532	5	384614

Figure 12 Industry Level Analysis

Industries Accommodation, Retail and Construction have ~50% of the total merchants in the data. Industries Accommodation and Retail have lower turnover per merchant. Similar is also true for educational services but it has only one merchant Healthcare and Finance & Insurance have quite high turnover per merchant.

Bank ID	# Months	# Credit Transactions	# Debit Transactions	Total Credit Amount	Total Debit Amount	Monthly Avg Credit Transactions	Monthly Avg Debit Transactions	Monthly Avg Credit Amount	Monthly Avg Debit Amount	# of Merchants	% Volume of Credits
8545	14	654	1331	11916982	11919504	47	95	851213	851393	2	43%
8544	17	1437	3866	7233807	7226247	85	227	425518	425073	4	26%
8534	16	5441	3311	3772067	3739199	340	207	235754	233700	5	13%
8535	14	1649	4665	2324373	2318574	118	333	166027	165612	6	8%
10656	13	57	849	1444859	1466503	4	65	111143	112808	1	5%
9966	11	276	1985	445920	432975	25	180	40538	39361	1	2%
6192	12	35	246	325133	346662	3	21	27094	28889	1	1%
9262	9	369	1694	231849	229561	41	188	25761	25507	1	1%
10591	2	503	86	209684	161625	252	43	104842	80813	1	1%
10479	9	121	454	105933	100592	13	50	11770	11177	1	0%

Figure 13 Bank ID Level Analysis

Bank IDs 8545 and 8544 have ~70% of the total transaction amount among all merchants. Bank ID 8545 has merchant 330698 which has a really high turnover. Bank IDs 8534 and 8535 have a significant presence among the merchants, ~50% of merchants have at least one account with these banks.

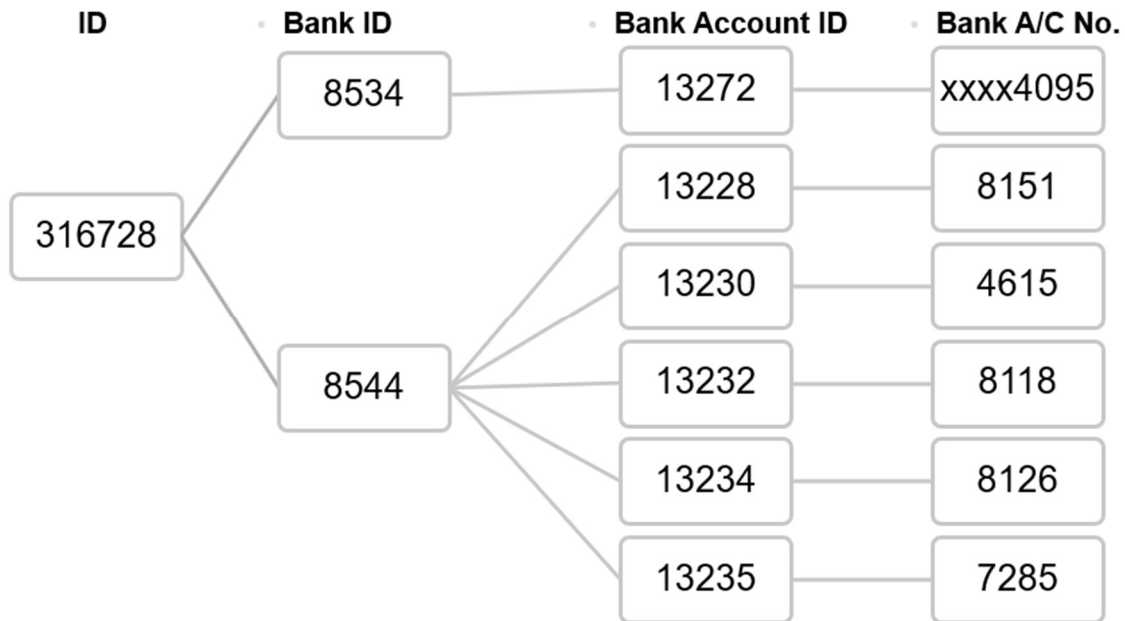


Figure 14 Visualization of Relationship

ID 316728 has total 6 accounts in 2 banks. Lead ID shows the Merchant, Bank ID represents the different banks in which merchant has accounts, bank account id is the account held by the merchant within that bank and account number is the account associated with that account id. This relationship is possible for all merchants but ID 321146 where the merchant has two bank account IDs within a bank but only one account number mapping for those ID's. Hence the one-to-one relationship between account ID and account number does not hold for that merchant.

ID	Bank ID	Bank Account ID	Account Number	Industry	Post Date	Description	Transaction Type	Amount
314036	9966	13226	xxxx7053	Retail Trade	4/27/2016	ELECTRONIC/ACH CREDIT TIMEPAYMENT CORP DIR DEP	credit	20900
314036	9966	13226	xxxx7053	Retail Trade	12/7/2016	WIRE TRANSFER INCOMING FEDWIRE CR TRN #xx5887	credit	19352
314036	9966	13226	xxxx7053	Retail Trade	10/11/2016	Credit with image ATM DEPOSIT	credit	12985
314036	9966	13226	xxxx7053	Retail Trade	11/14/2016	ELECTRONIC/ACH CREDIT KABBAGE INC LOAN	credit	12300
314036	9966	13226	xxxx7053	Retail Trade	9/16/2016	ELECTRONIC/ACH CREDIT LEASE CORP OF AM CASH Camped	credit	9999.99

Figure 15 Top 5 Deposits of Merchants

ID 314036: There is a loan granted by Kabbage in the recent history indicating higher risk as more debt is taken in the recent months.

ID	Bank ID	Bank Account ID	Account Number	Industry	Post Date	Description	Transaction Type	Amount
330698	8545	14374	1693	Health Care	4/18/2016	ACH deposit WEBBANK WBxxxx0415 PERSONA DOCTORS	credit	150000
330698	8545	14374	1693	Health Care	4/15/2016	ACH deposit KABBAGE INC LOAN x2138	credit	100000
330698	8545	14374	1693	Health Care	12/20/2016	ACH deposit BANKCARD BTOT DEP PERSONA DOCTORS-AUTHOR ...1496	credit	85492.04
330698	8545	14374	1693	Health Care	3/1/2016	ACH deposit BANKCARD MTOT DEP PERSONA DOCTORS-AUTHOR ...1496	credit	75345.79
330698	8545	14374	1693	Health Care	10/18/2016	ACH deposit BANKCARD BTOT DEP PERSONA DOCTORS-AUTHOR ...1496	credit	67437.08

Figure 16 Top 5 Deposits of Merchants part 2

ID 330698: There is a loan granted by Kabbage in April 2016 indicating higher risk as more debt is taken in the recent months.

ID	# Days Between Withdrawals
318465	0.15
316728	0.15
314036	0.16
321671	0.20
326062	0.22
325142	0.24
313082	0.26
330698	0.38
321218	0.40



ID	# Days Between Withdrawals
321356	0.43
326050	0.45
321380	0.46
321146	0.49
329803	0.54
323253	0.57
312745	0.61
308148	0.80
328212	1.01
325330	1.16
314559	1.19
310443	1.39

Figure 17 Number of Days between Withdrawals

We should look at number of days between credit transactions as well since it suggests the frequency in which the money is coming in the account and both these metrics in

conjunction can provide healthy information about the money churning in the account of the merchant.

The lower the number of days between withdrawals suggests that the merchant is paying its suppliers quite frequently. The highlighted merchants are making prompt payments to suppliers and have healthy cash flows in their business.

ID	COV
329803	189.7
310443	141.5
323253	130.6
314036	122.6
321218	120.6
314559	105.3
321356	101.4
325142	97.7
313082	95.6
308148	89.9

ID	COV
316728	87.9
328212	86.1
321380	77.1
318465	75.9
312745	69.6
326062	66.5
325330	64.6
326050	60.6
330698	51.6
321146	-280.4
321671	-427.8

Figure 18 Covariance Analysis

COV would help us to analyze the stability of the daily balances irrespective of the business size, lower the value of COV more stable are the daily balances of the merchants. The highlighted merchant ID's (left table) have a COV greater than 100 which means that their daily balances are varying a lot and are not that stable in nature

The merchant ID's with COV between 50 to 100 have relatively stable daily balances.

There are 2 merchant ID's (right table) with negative balances, their daily balances are also shifting a lot and not stable.

- ID - 308148

ID	Bank ID	Bank Account ID	Account Number	Industry	Post Date	Description	Transaction Type	Amount	Balance
308148	8535	12460	xxxx9928	Accommodation and Food Services	5/12/2016	AMAZON MKTPLACE PMTS AMZN.COM/BILL WA 05/12	debit	55.99	897.41
308148	8535	12460	xxxx9928	Accommodation and Food Services	5/13/2016	AMAZON MKTPLACE PMTS AMZN.COM/BILL WA 05/12	debit	117.96	690.12
308148	8535	12460	xxxx9928	Accommodation and Food Services	5/19/2016	AMAZON MKTPLACE PMTS AMZN.COM/BILL WA 05/18	credit	117.96	808.08
308148	8535	12460	xxxx9928	Accommodation and Food Services	5/20/2016	AMAZON MKTPLACE PMTS AMZN.COM/BILL WA 05/20	credit	55.99	864.07
308148	8535	12460	xxxx9928	Accommodation and Food Services	11/14/2016	AMAZON MKTPLACE PMTS AMZN.COM/BILL WA 11/12	debit	29.99	6326.34
308148	8535	12460	xxxx9928	Accommodation and Food Services	11/16/2016	AMAZON MKTPLACE PMTS AMZN.COM/BILL WA 11/16	debit	12.71	5586.57

- ID - 314559

ID	Bank ID	Bank Account ID	Account Number	Industry	Post Date	Description	Transaction Type	Amount	Balance
314559	8534	13271	xxxx3965	Information Technology	3/30/2016	xx2667 Palazzo E DES:WEB PMTS ID:Y35YR2 INDN:Egor Kovalev CO ID:1273	debit	3749.07	11526.02
314559	8534	13271	xxxx3965	Information Technology	10/6/2016	xx2667 Palazzo E DES:WEB PMTS ID:R1CD43 INDN:Egor Kovalev CO ID:1273	debit	350	18495.51
314559	8534	13271	xxxx3965	Information Technology	10/6/2016	xx2667 Palazzo E DES:WEB PMTS ID:S1CD43 INDN:Egor Kovalev CO ID:1273	debit	314.64	18180.87

Figure 19 Cluster of similar transactions

Above figure identified the group of similar transactions in our data.

### 3.11 Research Design Limitations

Several restrictions were discovered during the course of this research study. The primary topic of this study was so broad that creating a complete list of transaction keywords could have taken a year due to the SME knowledge required for the understanding of the keywords. Keywords were produced using text mining techniques. Keywords were 1-, 2-, and 3-gram word tokens retrieved from transaction descriptions, and verified using extensive knowledge of lending.

There was nothing significant found in the literature similar to the subject of this study, therefore almost everything in this study was done from scratch, which took a large amount of time. A thorough literature study was conducted on areas related to this research. Another drawback was that transaction descriptions did not use simple English and contained various finance and banking-related terminology and acronyms. Text pre-processing was not an easy or straightforward process. It was discovered that machine

learning models along with enhanced keywords list with such complicated data produced standard outcomes (83.56% accuracy) when compared to the rule-based engine. However, when the same model applied to unseen Market like India for stress testing, it was found that the accuracy significantly dropped to 66.72% indicating that the keywords applicable to the US Market could not be applied directly to the India Market. It means that for India market substantial data would be needed to generate new set of keywords, later these could be merged with the overall keywords master data sheet for generalization or the generic model. Large Language Models are also explored for performing the transactions classification. The methodologies employed in this study, as well as the findings, are particular to small and medium-sized firms in the United States and India Lending Markets.

The results acquired are determined by the quantity of the input data; if larger datasets were utilized, an exhaustive list of keywords would be generated. Machine learning models' performance should have improved in this situation, given that the training dataset was significantly larger than the one utilized in this study. Finally, the raw data lacked labels and classes, hence the main task in this study was to enhance these classes. It would have been ideal if keywords or classes had already been assigned to the transactions so that different supervised machine learning methods along with Generative AI LLM Models could have been tested and the relative performance of these models could have been measured in order to add more value to the study.

### **3.12 Conclusion**

This chapter covered data requirements, data extraction, and related problems. This portion covered data pre-processing for this study, as well as other text data transformation strategies. Two approaches to this study were described in detail, along with logical flow diagrams. This chapter additionally discusses in depth on methods and evaluation criteria. The tools and technology necessary to carry out this research were highlighted. Two approaches were used. One of them devised keywords, which were then used to classify transactions. In the other technique, we took categorized data and used a random forest



algorithm and large language model to do multi-class classification. At the end model is used with unseen data from different geography i.e. India Lending Market and performance recorded. Learnings are elaborated and improvements are recommended in later sections following this part of the study. Both methodologies were assessed separately, and in the post-classification process, important indicators were obtained. The Bank statement analyzer produced these indicators that helps underwriters to make credit decisions for SME loans.

When compared to the rule-based engine, machine learning models combined with an improved keywords list produced standard results (83.56% accuracy) with such complex data. However, when the same model was applied to an unknown market such as India for stress testing, the accuracy dropped significantly to 66.72%, demonstrating that keywords appropriate to the US market could not be applied straight to the India market. It indicates that for the India market, a large amount of data would be required to develop a new set of keywords, which could then be combined with the overall keywords master data sheet for generalization or the generic model. Large Language Models are also explored to perform transaction classification. Large Language Model were used with extensive prompt engineering which is a separate field in itself and it was found that LLM could perform much better as compared to Machine Learning Model as data is textual and it depends entirely on prompting strategy. However, LLM API model could be developed in future study to produce the classification of high dimensional complex data for producing an overall classification of entire dataset of transactions.

## CHAPTER IV:

### RESULTS

#### 4.1 Research Question One

The first research question as described in detail in chapter 1 & 3, was - To enhance the list of keywords used by the study (Mehta, 2020) for categorizing banking transactions of small and medium enterprises in the United States.

##### 4.1.1 Keywords Generation Process

The keyword creation process is the same as that used in the study (Mehta, 2020), which was used to produce keywords that would subsequently be utilized in the classification of banking transactions for small and medium-sized firms in the United States. Initially, the entire dataset was divided into two sets based on transaction type: credit or debit. Following pre-processing, transaction descriptions for credit and debit datasets were sent to the count vectorizer. This operation performed tokenization; n-grams (n = 1, 2, and 3) tokens were generated as a result of count vectorization, along with their frequencies of occurrence in the transaction description column. The results of debit and credit-related keywords were sorted in descending order of frequency and then combined to create a consolidated list of all keywords from transaction descriptions.

Keywords generated using the Text Mining method must be confirmed. Furthermore, these keywords were assigned to classes (First Level Classification), categories (Second Level Classification), and groups (High Level Classification). Based on lending domain experience and consultation with multiple underwriters, this list was reduced and refined to 1603 keywords in the study (Mehta, 2020); however, in the current study, this list has been increased to 3000. This is a comprehensive feature design initiative, and in order to improve the performance of the classification model developed by the study (Mehta, 2020), 3000 keywords were understood, verified, and mapped to higher levels of grouping.

The sample of keywords generated during the new study are placed below for reference and understanding. Each keyword is categorized into first, second and third or final level.

#	Keyword	First Level	Second Level	Third Level	Type
1	HOUSE LEASE	Loans	Debt Accrual	Liability	CREDIT
2	FEDWIRE CREDIT	Transfers	Inter-Account Transfers	Transfer	CREDIT
3	FLOORPLAN	Loans	Debt Accrual	Liability	CREDIT
4	BMW	Automotive Repair/Maintenance	Automobile Supplies/Services Revenues	Revenue	CREDIT
5	BANKCARD SETTLEMENT	Trade Payments	Trade Payments	Expenditure	DEBIT
6	BANKCARD SETTLEMENT	Trade Revenue	Trade Revenue	Revenue	CREDIT
7	BOB EVANS	Restaurants/Dining	Foodservice Payments	Expenditure	DEBIT
8	CHASE HOME LOAN	Loans	Mortgage	Liability	DEBIT
9	UTILITY BILL PAYMENT	Utilities	Utility Payments	Expenditure	DEBIT
10	IRS REFUND	Taxation	Tax Refund	Income	CREDIT

#	Keyword	First Level	Second Level	Third Level	Type
11	DIRECT DEPOSIT PAYROLL	Income	Salary	Employment Income	CREDIT
12	CAR PAYMENT	Loans	Auto Loan	Liability	DEBIT
13	ONLINE SHOPPING	Retail	E-commerce	Expenditure	DEBIT
14	GAS STATION	Transportation	Fuel	Expenditure	DEBIT
15	GROCERY STORE	Retail	Grocery	Expenditure	DEBIT
16	PHONE BILL	Utilities	Telecom Payments	Expenditure	DEBIT
17	CREDIT CARD PAYMENT	Debt Repayment	Credit Card Payment	Expenditure	DEBIT
18	RENT PAYMENT	Housing	Rent	Expenditure	DEBIT
19	HEALTH INSURANCE	Insurance	Health Insurance Premiums	Expenditure	DEBIT
20	DIVIDEND CREDIT	Investments	Dividend Income	Income	CREDIT
21	CAR INSURANCE	Insurance	Auto Insurance Premiums	Expenditure	DEBIT
22	ATM WITHDRAWAL	Cash Management	ATM Withdrawal	Expenditure	DEBIT
23	ATM DEPOSIT	Cash Management	ATM Deposit	Income	CREDIT

#	Keyword	First Level	Second Level	Third Level	Type
24	MORTGAGE PAYMENT	Loans	Mortgage	Liability	DEBIT
25	STOCK PURCHASE	Investments	Stock Purchase	Expenditure	DEBIT
26	STOCK SALE	Investments	Stock Sale	Income	CREDIT
27	BUSINESS EXPENSE	Business Expenses	Operational Costs	Expenditure	DEBIT
28	LOAN DISBURSEMENT	Loans	Loan Disbursement	Income	CREDIT
29	TAX PAYMENT	Taxation	Tax Payment	Expenditure	DEBIT
30	CASHBACK REWARD	Rewards	Cashback	Income	CREDIT
31	SERVICE FEE	Fees	Service Charges	Expenditure	DEBIT
32	SAVINGS INTEREST	Savings	Interest Income	Income	CREDIT
33	WAGE GARNISHMENT	Employment	Wage Garnishment	Expenditure	DEBIT
34	UTILITY REBATE	Utilities	Utility Refund	Income	CREDIT
35	LOAN REPAYMENT	Loans	Loan Repayment	Liability	DEBIT
36	SCHOLARSHIP CREDIT	Education	Scholarship	Income	CREDIT

#	Keyword	First Level	Second Level	Third Level	Type
37	STUDENT LOAN PAYMENT	Loans	Student Loan	Liability	DEBIT
38	CHARITY DONATION	Donations	Charity	Expenditure	DEBIT
39	MEDICAL BILL	Healthcare	Medical Expenses	Expenditure	DEBIT
40	HEALTHCARE REIMBURSEMENT	Healthcare	Insurance Reimbursement	Income	CREDIT
41	TRADE INCOME	Trade Revenue	Sales	Income	CREDIT
42	TRADE EXPENSE	Trade Payments	Purchase Costs	Expenditure	DEBIT
43	PROFIT SHARE	Investments	Profit Distribution	Income	CREDIT
44	TIPS RECEIVED	Employment	Tips	Income	CREDIT
45	RENTAL INCOME	Real Estate	Rental Income	Income	CREDIT
46	RENTAL PAYMENT	Real Estate	Rental Payment	Expenditure	DEBIT
47	PENSION PAYMENT	Retirement	Pension	Income	CREDIT
48	ALIMONY PAYMENT	Legal Obligations	Alimony	Expenditure	DEBIT
49	CHILD SUPPORT	Legal Obligations	Child Support	Expenditure	DEBIT

#	Keyword	First Level	Second Level	Third Level	Type
50	GOVERNMENT BENEFIT	Government Assistance	Benefits	Income	CREDIT
51	ONLINE TRANSFER	Transfers	Online Transfer	Transfer	DEBIT
52	MOBILE DEPOSIT	Deposits	Mobile Deposit	Income	CREDIT
53	WIRE TRANSFER IN	Transfers	Incoming Wire	Income	CREDIT
54	WIRE TRANSFER OUT	Transfers	Outgoing Wire	Transfer	DEBIT
55	INSURANCE CLAIM	Insurance	Claim Payout	Income	CREDIT
56	INSURANCE PREMIUM	Insurance	Premium Payment	Expenditure	DEBIT
57	HOTEL STAY	Travel	Accommodation	Expenditure	DEBIT
58	AIRLINE TICKET	Travel	Airfare	Expenditure	DEBIT
59	CAR RENTAL	Transportation	Car Rental	Expenditure	DEBIT
60	MEAL REIMBURSEMENT	Travel	Reimbursement	Income	CREDIT
61	LOST ITEM REFUND	Retail	Refund	Income	CREDIT
62	BOOK PURCHASE	Retail	Bookstore	Expenditure	DEBIT

#	Keyword	First Level	Second Level	Third Level	Type
63	TECH SUPPORT	Services	Technical Support	Expenditure	DEBIT
64	REFUND RECEIVED	Refunds	Retail Refund	Income	CREDIT
65	OVERDRAFT FEE	Fees	Overdraft	Expenditure	DEBIT
66	CREDIT LIMIT INCREASE	Loans	Credit Line	Income	CREDIT
67	CREDIT LIMIT DECREASE	Loans	Credit Line	Expenditure	DEBIT
68	GAS BILL	Utilities	Gas Payment	Expenditure	DEBIT
69	WATER BILL	Utilities	Water Payment	Expenditure	DEBIT
70	LEGAL FEES	Legal Expenses	Attorney Fees	Expenditure	DEBIT
71	INVESTMENT INCOME	Investments	Interest	Income	CREDIT
72	TRAVEL EXPENSE	Travel	Travel Costs	Expenditure	DEBIT
73	FUEL REIMBURSEMENT	Transportation	Fuel Reimbursement	Income	CREDIT
74	UNPAID ITEM	Banking	Returned Check	Liability	DEBIT
75	FOREIGN EXCHANGE FEE	Fees	Currency Exchange	Expenditure	DEBIT



#	Keyword	First Level	Second Level	Third Level	Type
76	REMITTANCE RECEIVED	Transfers	Remittance	Income	CREDIT
77	REMITTANCE SENT	Transfers	Remittance	Transfer	DEBIT
78	MEMBERSHIP FEE	Services	Membership	Expenditure	DEBIT
79	ENTERTAINMENT EXPENSE	Entertainment	Recreational Spending	Expenditure	DEBIT
80	ROYALTY PAYMENT	Intellectual Property	Royalties	Income	CREDIT
81	FREELANCE PAYMENT	Employment	Freelance	Income	CREDIT
82	SUBSCRIPTION FEE	Services	Subscription	Expenditure	DEBIT
83	SOCIAL SECURITY INCOME	Government Benefits	Social Security	Income	CREDIT
84	STUDENT GRANT	Education	Grants	Income	CREDIT
85	FOREIGN TRANSFER	Transfers	International Transfer	Transfer	DEBIT
86	MOBILE PAYMENT	Transfers	Mobile Payment	Transfer	DEBIT

#	Keyword	First Level	Second Level	Third Level	Type
87	CASH WITHDRAWAL	Cash Management	Withdrawal	Expenditure	DEBIT
88	MERCHANT PAYMENT	Trade Payments	Vendor Payment	Expenditure	DEBIT
89	MERCHANT CREDIT	Trade Revenue	Vendor Credit	Income	CREDIT
90	EDUCATION EXPENSE	Education	Tuition	Expenditure	DEBIT
91	REWARDS POINTS	Rewards	Points Earned	Income	CREDIT
92	BONUS PAYMENT	Employment	Bonus	Income	CREDIT
93	TUITION PAYMENT	Education	Tuition	Expenditure	DEBIT
94	PRODUCT RETURN REFUND	Retail	Refund	Income	CREDIT
95	CASH BACK REWARDS	Rewards	Cashback	Income	CREDIT
96	RENTAL DEPOSIT	Real Estate	Security Deposit	Income	CREDIT
97	RENTAL DEPOSIT REFUND	Real Estate	Security Deposit Refund	Income	CREDIT
98	CREDIT REVERSAL	Banking	Payment Reversal	Income	CREDIT

#	Keyword	First Level	Second Level	Third Level	Type
99	DEBIT REVERSAL	Banking	Payment Reversal	Expenditure	DEBIT
100	MONEY ORDER	Banking	Money Order Purchase	Expenditure	DEBIT

Figure 20 Sample Keyword List

There were 45 level 2 categories to which 74 level 1 categories were mapped to. For example – Money order of category 1 “Banking” is mapped to category level 2 “Money Order Purchase” and finally to the Expenditure category. There is final 5 categories based on which the cash flow variables were generated. These are Revenue, Expenditure, Transfers, Debt Service and Liabilities. The list of 3000 keywords produced and used for classification as a part of this study is attached in the appendix for the reference.

#### 4.1.2 Keywords validations

This research produced keywords, which are 1,2, and 3 grams of word tokens present in the parsed transaction descriptions. Keywords that appeared frequently were considered, and the total number was enhanced to 3000. These 3000 keywords were validated by underwriters and key lending subject matter specialists. These 3000 keywords were also assigned to classes and groups and entered into a keywords file.

#### 4.2 Research Question Two

To enhance the performance of the machine learning model, build in the study (Mehta, 2020) for categorizing banking transactions of small and medium enterprises in the United States.

#### **4.2.1 Machine Learning Classification**

Random Forest Classifier was built and model was created from processed training data of 70,000 banking transactions. Default parameters were used except the number of trees. The number of trees was set to 450. Entropy was used as a method for splitting on the node. The input data is labelled using clustering and rule-based classification. DBSCAN is used the clustering mechanism for grouping similar transactions and rule-based engine is used to assign the label to each transaction using the keywords. This process is same as what explained in the study (Mehta, 2020). The model used is also same as what was used in the study (Mehta, 2020) but the list of keywords is different and more i.e. 3000 as compared to 1603 used in the study (Mehta, 2020). The accuracy of the classification achieved in the model developed in the study (Mehta, 2020) was 73.62% whereas the accuracy has been increased to 83.56% by just doubling the number of keywords pertaining to the banking transactions.

#### **4.2.2 Machine Learning based classification Evaluation**

The Random Forest model was run using 30% test data (30,000 banking transactions), and the model-assigned or predicted keywords were compared to the actual keywords. Thus, accuracy was the measure utilized to evaluate the model, and the accuracy reached after running the random forest model was 83.56%. The post-classification module was simply a calculation module for various cash flow variables. The derived variables for nine clients were manually examined and verified with formulas.

### **4.3 Research Question Three**

The third research question in this study was “To perform the classification of banking transactions of SME into the above identified categories using novel AI that includes Large Language Models.”

### 4.3.1 LLM Classification Results

Large Language Models (LLMs) such as ChatGPT or those accessible via native APIs are effective tools for multi-class classification tasks. These models are pre-trained on massive volumes of text data, allowing them to comprehend and produce human-like language. For example, given an input phrase (such as a transaction description), a prompt can be created that instructs the model to select the most appropriate category from a predefined list of alternatives. Using its grasp of language and context, the model may then construct the category that best matches the input.

Prompt engineering is critical for successfully using LLMs for multi-class classification. Prompt engineering is the process of designing input prompts in such a way that the model's output is as accurate as possible. For example, to promote accurate categorization, prompts could include precise instructions, contextual examples, or a clear structure that directs the model to the correct response. In multi-class classification, well-designed prompts can assist the model in distinguishing between similar categories and producing consistent results. This approach is especially crucial because LLMs are very sensitive to the wording and structure of the prompts they get.

Using LLMs for multi-class classification has various advantages over conventional machine learning models such as Random Forests. One of the primary advantages is LLMs' ability to generalize across multiple tasks without requiring considerable retraining. While classic models, like as Random Forests, require structured input data and frequently a large number of labeled examples for training, LLMs can do zero-shot or few-shot learning, which allows them to classify new cases with little to no extra training. Furthermore, LLMs grasp language context and nuances, resulting in more accurate and contextually appropriate classifications, particularly in case of the complex data. In contrast, models such as Random Forests may struggle with tasks requiring significant semantic understanding in the absence of extensive feature engineering. Furthermore, LLMs are a more flexible and scalable solution since they can be easily transformed to new tasks by simply changing the prompts, whereas traditional models require retraining or fine-tuning

to respond to new data. Because of these reasons, it was worth exploring the LLM for this study. Next sub section will describe the results of using LLM model for this study to classify the banking transactions.

#### 4.3.1.1 PROMPT TEMPLATES AND LLM RESPONSES

To classify the banking transaction descriptions into the categories of Revenue, Debt Service, Liability, Transfers, and Expenditure, we can create a variety of prompt templates. These templates are designed to guide the model toward accurately categorizing the transactions based on the information provided. Below are several different prompt designs used in this study to achieve this goal.

Direct Classification Prompt:

Classify the following banking transaction into one of the categories: Revenue, Debt Service, Liability, Transfers, or Expenditure. Transaction description: “ATM CASH DEPOSIT ON 11/30 10 CHESTNUT AVE SO. SAN FRAN CA XXX8084 ATM ID 0509A CARD 9367”.

LLM Response –

The transaction description "ATM CASH DEPOSIT ON 11/30 10 CHESTNUT AVE SO. SAN FRAN CA XXX8084 ATM ID 0509A CARD 9367" would be classified as Revenue. This classification is based on the nature of the transaction, which involves a cash deposit into an account. And this is the correct classification based on the rule-based engine and SME knowledge.

Contextual Understanding Prompt:

Based on the context of typical banking transactions, categorize the following description into Revenue, Debt Service, Liability, Transfers, or Expenditure: “CREDIT FOR MERCHANT BNKCD DEPOSIT CO REF- xxxxxxxx2883”.

LLM Response - The transaction description "CREDIT FOR MERCHANT BNKCD DEPOSIT CO REF- xxxxxxxx2883" would be categorized as Revenue.

This classification is based on the context that the credit for a merchant deposit typically represents income received, which falls under the Revenue category. And this is the correct classification based on the rule-based engine and SME knowledge.

Role-Based Prompt:

As a banking expert, determine the correct category for this banking transaction description. Choose from Revenue, Debt Service, Liability, Transfers, or Expenditure.

Transaction: "EMI DEBIT EBF PARTNERS LLC xxxxxxxxxxxx9279".

LLM Response to above is –

The transaction description "EMI DEBIT EBF PARTNERS LLC xxxxxxxxxxxx9279" would be categorized as Debt Service.

This classification is based on the fact that an EMI (Equated Monthly Installment) debit typically represents a regular payment toward a loan or debt, which is a characteristic of Debt Service. And this is the correct classification based on the rule-based engine and SME knowledge.

Instructional Prompt:

Please categorize the following transaction into the most appropriate category: Revenue, Debt Service, Liability, Transfers, or Expenditure. The transaction is described as: "PURCHASE AUTHORIZED ON 01/19 SHELL Service Station SCOTTSDALE AZ Pxxxxxxxxxxxx3556 CARD 8420".

LLM Response –

The transaction description "PURCHASE AUTHORIZED ON 01/19 SHELL Service Station SCOTTSDALE AZ Pxxxxxxxxxxxx3556 CARD 8420" would be categorized as Expenditure.

This classification is based on the fact that a purchase at a service station typically involves spending money, which falls under the Expenditure category. And this is the correct classification based on the rule-based engine and SME knowledge.

#### Scenario-Based Prompt:

Imagine you are reviewing a list of transactions for credit decisioning. How would you classify the following transaction? Choose between Revenue, Debt Service, Liability, Transfers, or Expenditure: “Payment from loanagency1 Data Services”.

#### LLM Response –

In the context of reviewing transactions for credit decisioning, the transaction description "Payment from loanagency1 Data Services" would be classified as Revenue.

This is because a payment from a loan agency typically represents income received, which falls under the Revenue category. Interestingly LLM model used the context knowledge and determined the category correctly as this is the payment received from loan agency whereas the ML model classified this into Liability based on the keyword “Loanagency”.

#### Example-Based Prompt:

Consider the transaction description below. Based on your understanding of similar transactions, classify it as either Revenue, Debt Service, Liability, Transfers, or Expenditure: “MONEY TRANSFER AUTHORIZED ON 07/31 FROM Uber Uber CA PXXXXXXXXXXXXXXXXX7555 CARD 8749”.

LLM Response - The transaction description "MONEY TRANSFER AUTHORIZED ON 07/31 FROM Uber Uber CA PXXXXXXXXXXXXXXXXX7555 CARD 8749" would be classified as Transfers.

This classification is based on the fact that it involves a transfer of money between accounts, which fits the Transfers category. And this is the correct classification based on the rule-based engine and SME knowledge.

#### Comparative Analysis Prompt:

Given the banking transaction below, how would you classify it in comparison to other transactions. Choose the best fit from Revenue, Debt Service, Liability, Transfers, or Expenditure. Transaction: “Withdraw from Credit Card”

#### LLM Response –



The transaction description "Withdraw from Credit Card" would be classified as Debt Service.

This classification is based on the fact that withdrawing from a credit card typically involves accessing credit and thus implies a potential obligation to repay, which aligns with the Debt Service category.

When the same prompt used again the output received was different and also correct – The transaction “Withdraw from Credit Card” would best fit into the Liability category. This is because withdrawing from a credit card increases the amount owed on the credit card, which represents a liability for the account holder. However, the ML classification model based on the keywords has marked this transaction as Revenue which is off course an incorrect decision as this is a withdraw from the credit card and not from the account.

Analytical Prompt:

Analyze the following transaction description and categorize it into one of the following: Revenue, Debt Service, Liability, Transfers, or Expenditure. Description: “Funding Transfer Credit FROM ...9165”

LLM Response –

The transaction description "Funding Transfer Credit FROM ...9165" would be classified as Transfers.

This classification is based on the fact that the description indicates a credit related to a transfer of funds, which falls under the Transfers category.

When the same prompt used again the output received was the same but the description was bit more elaborated.

The transaction described as “Funding Transfer Credit FROM ...9165” would best fit into the Transfers category.

This is because the description indicates that funds are being transferred from one account (ending in 9165) to another, which typically represents an internal movement of funds

rather than an expense, revenue, or liability. On the other hand the actual output was Liability because the transfer is related to the credit from an account.

#### Categorization Challenge Prompt:

You have a challenging task to categorize various banking transactions. For the transaction described below, what category would you assign? Options: Revenue, Debt Service, Liability, Transfers, or Expenditure. Transaction:” Tfr from \*\*\*\*2611 to \*\*\*\*6927 House Budget Funding”

#### LLM Response –

The transaction description "Tfr from \*\*\*\*2611 to \*\*\*\*6927 House Budget Funding" would be classified as Transfers.

This classification is based on the fact that "Tfr" (transfer) indicates the movement of funds between accounts, which fits the Transfers category. However the ML classification model trained on keywords classified this transaction as Liability because of the appearance of the keyword called Funding. The actual output should be a transfer because even though this transaction is related to the House Budget Funding but it is being transferred from one account to another.

#### Financial Reporting Prompt:

For the purpose of accurate financial reporting, classify the following transaction into one of these categories: Revenue, Debt Service, Liability, Transfers, or Expenditure. Description: “FORTE - FUNDING ACH-X617-C458B”

#### LLM Response -

The transaction description "FORTE - FUNDING ACH-X617-C458B" would be classified as Transfers.

This classification is based on the term "FUNDING ACH" which typically refers to an Automated Clearing House (ACH) transfer, indicating the movement of funds between accounts. When the same prompt is used again the output received was different but

accurate - The transaction description "FORTE - FUNDING ACH-X617-C458B" suggests that it involves the receipt of funds, possibly through an Automated Clearing House (ACH) transfer. Since this appears to be incoming funds, the appropriate category for this transaction would likely be Revenue.

Prompt with Examples:

Here's a transaction description. Given examples of similar transactions, classify this one into the appropriate category (Revenue, Debt Service, Liability, Transfers, or Expenditure): "Chase Acct Collections XXX-XXX-0433".

Few Shot Examples -

Transaction "FORTE - FUNDING ACH-X617-C458B" is classified as Transfers.

Transaction "CAPITAL ONE CRCARDPMT xxxxxxxxxxxx7017 9xxxxx1285MURTHY KRISH" is classified as Debt Service.

LLM Response –

The transaction description "Chase Acct Collections XXX-XXX-0433" would be classified as Liability.

This classification is based on the context of "Collections," which typically involves the collection of overdue amounts, indicating an outstanding obligation or liability related to the account. This is the correct outcome.

Decision-Making Prompt:

Make a decision on the following transaction description. Into which category does it fall? Choose from Revenue, Debt Service, Liability, Transfers, or Expenditure:" AMERICAN EXPRESS COLLECTION XXXXXX1567 CCD ID: XXXXXX2250".

LLM Response –

he transaction description "AMERICAN EXPRESS COLLECTION XXXXXX1567 CCD ID: XXXXXX2250" would be classified as Debt Service.

This classification is based on the term "COLLECTION," which suggests a payment related to settling a debt or credit card obligation, aligning with the Debt Service category.

The transaction described as "AMERICAN EXPRESS COLLECTION XXXXXX1567 CCD ID: XXXXXX2250" appears to be related to the payment of a credit card bill or a collection on a credit card account. Given that this involves paying off a debt or a liability associated with a credit card, the appropriate category for this transaction would be Debt Service. The same transaction was classified as 'Liability' by the ML model which is a correct outcome.

Clarification Prompt:

To ensure clarity in credit decisioning process within the bank, categorize this transaction. Options: Revenue, Debt Service, Liability, Transfers, or Expenditure. Transaction: "CAPITAL ONE CRCARDPMT xxxxxxxxxxxx7017 9xxxxx1285MURTHY KRISH".

LLM Response –

The transaction description "CAPITAL ONE CRCARDPMT xxxxxxxxxxxx7017 9xxxxx1285MURTHY KRISH" would be classified as Debt Service.

This classification is based on "CRCARDPMT," which suggests a payment related to a credit card, indicating a debt repayment, which falls under Debt Service. This is a correct outcome.

Task-Oriented Prompt:

You have been given the task to classify various banking transactions. For the following transaction, select the correct category: Revenue, Debt Service, Liability, Transfers, or Expenditure. Description: "Wire Transfer Fee", "Cheque xxxxxxx2315", "Overdraft Protection Transfer Fee".

LLM Response –

Here's the classification for each of the provided transactions:

Wire Transfer Fee – Transfers. This is a fee related to the processing of a wire transfer, which involves moving funds between accounts.

Cheque xxxxxxx2315 – Transfers. This entry likely refers to a cheque that has been processed, representing a movement of funds between accounts.

Overdraft Protection Transfer Fee – Transfers. This fee is related to transferring funds to cover an overdraft, which involves moving funds to prevent an account from going negative.

Another time the same prompt was used and the different output is received which is correct also. This is because of the indeterministic nature of the generative AI –

The correct category for this transaction would be Expenditure.

This is because fees such as "Wire Transfer Fee" and "Overdraft Protection Transfer Fee" represent costs incurred by the account holder, and "Cheque xxxxxxxx2315" likely refers to a payment, which also falls under expenditure.

Based on the rule-based classification or SME review expenditure is the correct category for the above transactions as these are the expenses incurred by the SME consumer.

The sample of 10,000 banking transactions descriptions are considered for the classification using the LLM model due to the token limit associated with the LLM and time constraint. Out of 10,000 banking transactions 8129 transactions are correctly classified using LLM which clearly indicates that LLM is a good option for transactions classification that has required no training on the dataset. However, the missed classifications are again provided to the LLM and out of 1,871 missed classifications or incorrect classifications, 391 are classified into the correct category, this behavior is due to the non-deterministic nature of the generative AI. After clubbing the right classifications together –  $8129 + 391 = 8520$  transactions were classified correctly which is definitely a good indicator of the performance of the LLM Models. Thus, at the end we could say that the performance achieved with the LLM model was 85.20% after few shots learning.

The final template used for the prompting LLM is placed below for the reference –

Persona – The role that LLM has to assume

Instructions - The instructions described for LLM to perform

To do and not to do guidance – what is expected from the LLM is indicated here

Data Context – The data to be referred is placed here

Few Shot Examples – the examples to be referred during classification are placed here

The detailed template for one of the banking transactions is placed below –

Assume you are an analyst working in a bank and your job is to categorize the various banking transactions into the below categories –

Revenue - Revenue is the total amount of income earned by a small and medium enterprise through its business activities or operations. This comprises sales of goods and services, interest, dividends, commission on services provided, interests on investments and any other type of revenue.

Expenditure - Expenditure is the spending or outflow of funds by a small and medium enterprise. It is the entire amount spent on purchasing goods and services, paying personnel, shopping, paying taxes, paying commissions, fees and fulfilling obligations.

Debt service - Debt service is the cash required to cover the repayment of a debt's interest and principal over a specific time period. This covers payments on loans, bonds, and other types of borrowed funds.

Liability - Liability refers to a small and medium enterprises legal financial commitments or responsibilities incurred throughout the course of commercial activities. Liabilities are resolved over time by transferring economic rewards like money, products, or services.

Transfers - Transfers are the movements of funds from one account to another within the same organization or between separate organizations. Transfers could occur between different accounts of same or different customers.

You are instructed to use context knowledge of financial services domain including the information given above for the various classification categories.

Transactions descriptions to be classified are placed below for the reference –

MONTHLY SERVICE FEE Monthly Service Fee Summary open dialog

Wire Transfer Fee

External transfer fee - Next Day - 04/03/2018 Confirmation: xxxxx1512

ATM WITHDRAWALXX3254 07/X6360 E 72N

AMAZON MKTPLACE PMTS AMZN.COM/BILL WA 10/18

-

-

-

-

-

-

Few Shots Examples are placed below for your reference –

ONLINE BANKING TRANSFER TO ACCT \*9336 – Expenditure

MRCHNT PMNT PROC Deposit xxxxxxxx9032 CCD ID: WFBSPTACH2 – Revenue

CHASE CREDIT CRD EPAY XXXXXX XXXXXX4070 MARAM S KHALIL – Debt

Service

Funding Transfer Credit FROM ...9165 - Liability

TRANSFER FROM CHECKING \*\*\*\*\*xxxx xx-x0-16 BB&T M-APP

TRANSFER – Transfer

-

-

- 
- 
- 

### 4.3.2 Post Classification Results

There is no change in this portion of the study (Mehta, 2020) in a way how the output variables (cash flow variables) are extracted or generated but of course the values of the output variables obtained differs because of the changes in the modelling procedures by employing enhanced list of keywords using traditional AI classification that includes random forest classifier and at the same time LLM (Novel AI) classifier which is a niche technology at the moment.

### 4.3.3 Output Variables

These cashflow output variables are generated post the use of LLM Classifier. An example is given below for understanding.

Cust Id	502728	Cust Id	373338
Number of Accounts	47	Number of Accounts	13
Number of Bank Statements	122	Number of Bank Statements	111
Revenue	51645.56	Revenue	15408.28
Operating Costs	25442.57	Operating Costs	12670.23
Number of NSF's	3	Number of NSF's	1
Operating Profit	26202.99	Operating Profit	2738.045
Debt Repayment	0	Debt Repayment	3333.33
Total Credits	58220.56	Total Credits	16734.06
Total Debits	25487.84	Total Debits	6482.941
Debt Service	0	Debt Service	3333.33
DSCR	-999	DSCR	0.821414
Operating Profit Margin	0.507362	Operating Profit Margin	0.1777
Debt Repayment per Revenue	0	Debt Repayment per Revenue	0.216334



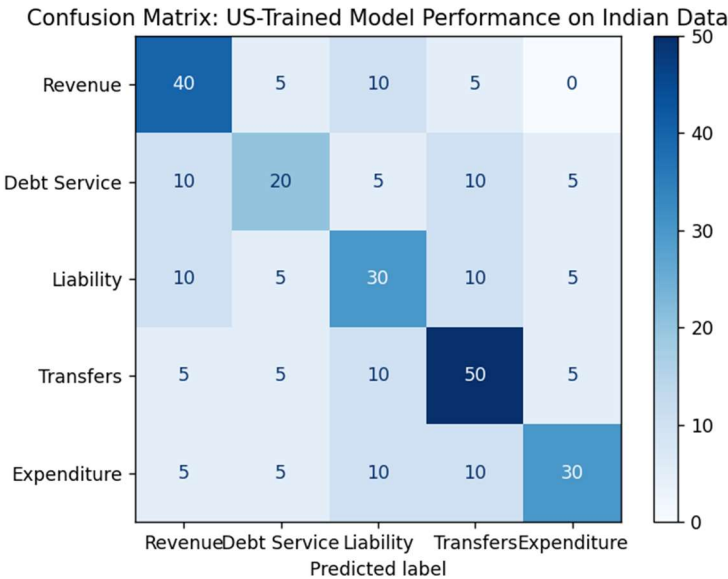
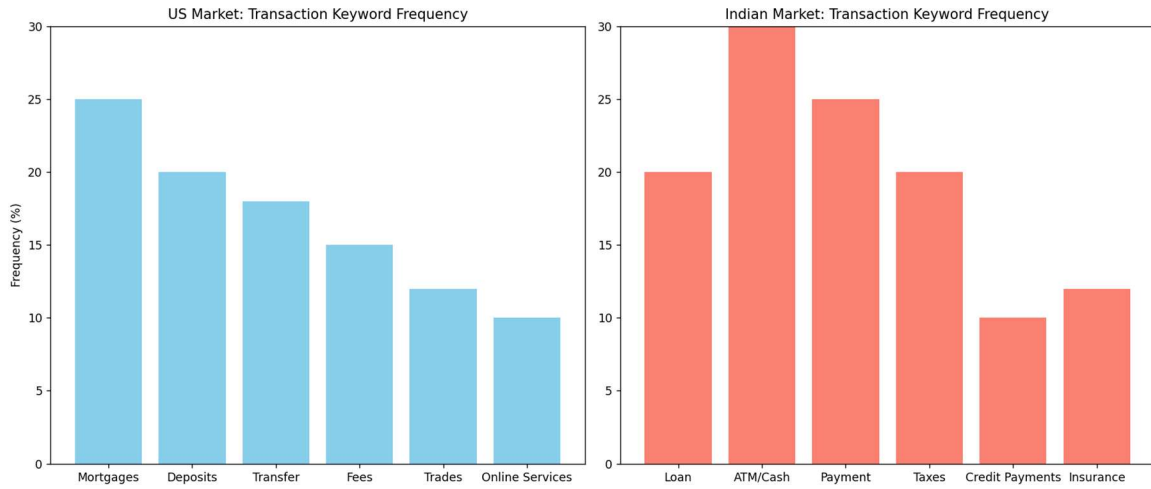
Debits to Income	0.437781	Debits to Income	0.38741
Cash Flow Ratio	0.462219	Cash Flow Ratio	0.51259

Figure 21 Set of output variables – LLM Model

As shown in the above table, for the customer with customer id 502728, below are the values for output variables. This customer with customer id 502728, has submitted 122 bank statements. Revenue for this customer was \$51645.56 and operating cost was calculated as \$25442.57. The operating profit for this customer was \$26202.99. The total Debt repayment amount was \$0. Debt Service amount was \$0 which is the same as the debt repayment amount indicating that customer does not have any debt liability. DSCR value was -999. DSCR i.e. ratio of operating profit to the debt service value signifies that since there is no debt taken by this customer the default value of DSCR is -999. Thus, based on these variables, the customer would be able to repay his debt if granted a new loan. The bad thing about this customer is that he appeared 3 times in the bracket of non-sufficient funds. The operating profit margin is also very good and here this is equal to 0. 507362. Debt to income ratio 0. 437781 is not very high and suitable for granting any new loans to the customer. It indicates that there were debits in his account equivalent to 0. 437781 times the credits. The cash flow ratio is considerable i.e. 0. 462219 indicating that the customer has 0. 462219 times the revenue amount cash available for paying off his debts, meeting his day-to-day requirements. Thus, based on the above values and interpretation, this customer will get a new loan funded by the fintech start-ups.

#### 4.4 Research Question Four

Last research question related to this study is “To apply the above study for different geographical location i.e. India lending Market to see if the same model has worked well for another location so that it could be generalized.”.

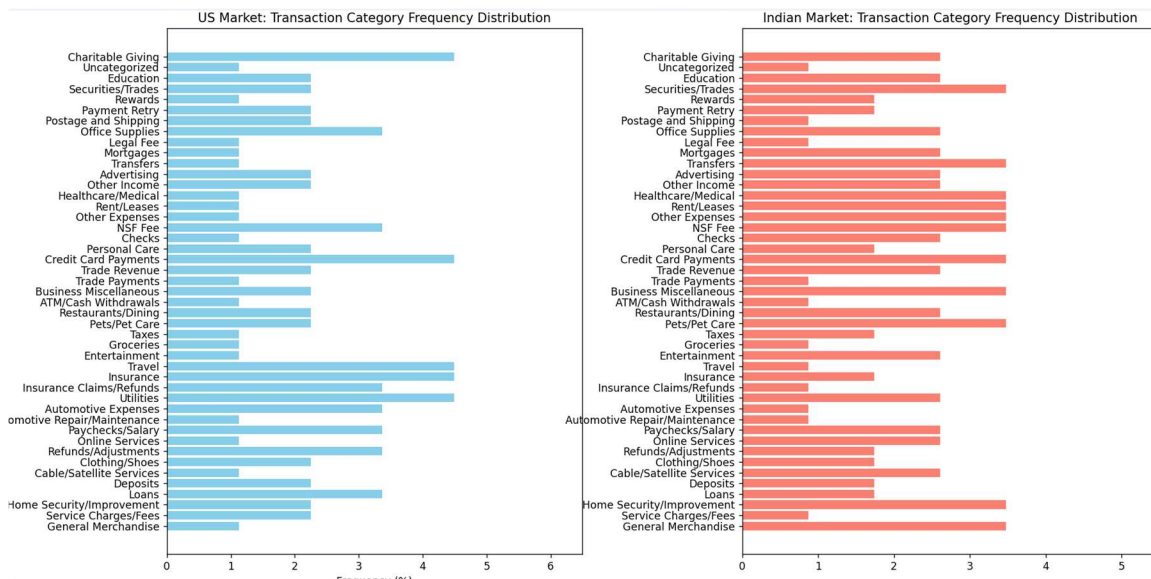


Frequency Distribution of Transaction Keywords: The first graph compares the frequency of common transaction keywords in the US and Indian markets. The US and India markets have different set of keywords with different frequencies. The graphs are plotted against first level of categorization considering few of them for better visibility.

Final Class/Category Distribution: The second graph shows the distribution of transactions across the five cash flow categories considered in the current study in both

the US and Indian markets. The US market has a more balanced distribution, while the Indian market shows a higher percentage in "Transfers" and "Liability" categories, reflecting regional banking practices.

Confusion Matrix for Multi Class Classification Model Performance on Indian Data: The third graph illustrates the performance of the US-trained classification model when applied to Indian data. The confusion matrix shows a significant number of misclassifications, particularly in categories like "Debt Service" and "Liability", indicating that the model struggles to accurately classify Indian transactions.



In the above figure this is clear that same set of keywords have very different frequency distributions across United States and India Lending Markets.

When compared to the rule-based engine, machine learning models combined with an improved keywords list produced standard results (83.56% accuracy) with such complex data. However, when the same model was applied to an unknown market such as India for stress testing, the accuracy dropped significantly to 66.72%, demonstrating that keywords appropriate to the US market could not be applied straight to the India market. It indicates that for the India market, a large amount of data would be required to develop

a new set of keywords, which could then be combined with the overall keywords master data sheet for generalization or the generic model.

#### **4.5 Summary of Findings**

As explained in sections above there were 4 research questions pertaining to this study. This study has been able to achieve all the 4 objectives. The keywords are enhanced and the performance of the model developed as a part of study (Mehta, 2020) was improved from 73.62% to 83.56% when the list of keywords is enhanced from 1603 to 3000. Moreover, the LLM classification model has shown significant hope and was explored with small population of the bank transactions limiting to 10,000 and accuracy achieved was 85.20%. Post classification model remains same for both traditional and LLM model and output variables are computed. Lastly, the classification model has applied to India Lending Market and the performance was found to significantly decrease to 66.72% indicating that US keywords are not applicable to India Market due to various regional reasons.

## CHAPTER V:

### DISCUSSION

#### **5.1 Discussion of Results**

The study demonstrates that enhancing the list of keywords from 1603 to 3000 improved the model's performance from 73.62% to 83.56%. This suggests that a more extensive and refined keyword list can significantly contribute to the accuracy and effectiveness of the classification model. The LLM classification model showed promising results with an accuracy of 85.20% when tested on a limited dataset of 10,000 bank transactions. This indicates that LLMs can effectively handle classification tasks even with a relatively small dataset. The study notes that the post-classification model remains consistent across both traditional and LLM models. This suggests that the downstream processes are stable and can handle outputs from different classification models without significant modifications. When the classification model was applied to the Indian lending market, performance dropped to 66.72%. This indicates that the US-based keyword list and model may not be fully applicable to the Indian context due to regional differences.

#### **5.2 Discussion of Research Question One**

The initial research question included extracting and enhancing keywords from bank transactions in order to conduct banking transaction analysis. Keywords are unigrams, bigrams, and trigrams derived from transaction descriptions. According to this study, it was able to produce keywords from parsed bank statements using text mining tools such as count vectorizers and n-gram methods; however, it was later shown that this methodology has limitations. The text was not in plain English and comprised a variety of finance terms, abbreviations, and banking terminology, thus just selecting the most often occurring tokens as keywords would not have worked for the Bank statement analyzer in this study. To address this constraint, this study employed lending domain experience to drill down into the keywords generated. Keywords list is enhanced from 1603 to 3000 in the current study and that has improved the performance of the classification model.

### **5.3 Discussion of Research Question Two**

A Random Forest Classifier was developed using 70,000 processed banking transactions as training data. The model's parameters were kept at their defaults, except for the number of trees, which was set to 450. Entropy was chosen as the method for node splitting. The input data was labelled using a combination of clustering and rule-based classification. DBSCAN was employed for clustering similar transactions, while a rule-based engine assigned labels to transactions based on keywords. This approach mirrors that of a previous study, with the main difference being the use of 3,000 keywords instead of 1,603. As a result, the accuracy of the classification improved from 73.62% in the previous study to 83.56%.

### **5.4 Discussion of Research Question Three**

The classification of 10,000 banking transaction descriptions was performed using a Large Language Model (LLM). This choice was influenced by the token limit associated with LLMs. Out of the 10,000 transactions, the LLM successfully classified 8,129 transactions correctly on the first attempt. This high accuracy rate demonstrates that LLMs are effective for classifying banking transactions, even without requiring prior training on the specific dataset.

Around 1,871 transactions were misclassified or not classified correctly by the LLM. To improve accuracy, these missed classifications were resubmitted to the LLM for another attempt at classification. On the second pass, the LLM correctly classified an additional 391 transactions. This improvement in accuracy on the second attempt can be attributed to the non-deterministic nature of generative AI, where different outputs can be produced from the same input due to the probabilistic approach of these models.

When combining the initial correct classifications (8,129) with the additional correct classifications from the second attempt (391), a total of 8,520 transactions were classified

correctly. This result suggests that the LLM model is quite effective in classifying banking transactions, with an overall performance accuracy of 85.20%.

The concept of "few-shot learning" plays a significant role in this process, where the LLM improves its accuracy after being exposed to a small amount of additional input data or examples. This iterative process allows the model to refine its classifications without requiring extensive retraining, demonstrating the adaptability and robustness of LLMs in handling real-world data classification tasks. The ability to achieve such a high level of accuracy without extensive training on the dataset indicates the potential of LLMs for similar classification tasks in various domains.

The use of an LLM model for classifying banking transaction descriptions proved to be highly effective, with an overall accuracy rate of 85.20% after a few-shot learning process. The LLM's performance underscores its capability to handle complex classification tasks with minimal prior training, making it a valuable tool in scenarios where time and resource constraints are critical. The improvement seen after reclassifying the missed transactions highlights the model's non-deterministic nature and its potential for further refinement and accuracy enhancement. This case study demonstrates the feasibility of LLMs in financial domain and other similar applications, where quick and accurate classification is essential.

## **5.5 Discussion of Research Question Four**

The analysis of transaction data from the US and Indian markets reveals significant differences in keyword frequency, category distribution, and model performance. The study includes several graphs and key findings that highlight the distinct characteristics of these two financial markets and the challenges of applying machine learning models across different regions. The study compares the frequency of common transaction keywords in the US and Indian markets. It is evident that the two markets have distinct sets of keywords, each with varying frequencies. The study focuses on the different levels of categorization, selecting a few keywords for better visibility. This comparison provides the versatility of

transaction behaviours in these two markets. For instance, a keyword that is highly frequent in the US market might be less common or even rare in the Indian market, and vice versa. This difference in keyword distribution reflects the varied consumer behaviour, transaction types, and financial practices in these regions.

Such a disparity in keyword frequency is crucial for this study, especially because this is based on keyword-driven classification. When models are trained on data from one market, like the US, they may not perform well in another market, such as India, because the keyword distribution, which heavily influences the model's decisions, is not consistent across regions. This study provides insights into the distribution of transactions across five cash flow categories: "Income," "Expense," "Transfers," "Debt Service," and "Liability." This distribution is analysed for both the US and Indian markets. The findings reveal that the US market has a more balanced distribution of transactions across these categories. In contrast, the Indian market shows a skewed distribution, with a higher percentage of transactions falling under the "Transfers" and "Liability" categories. This imbalance could be indicative of regional banking practices or consumer behaviour, where certain types of transactions are more prevalent.

Understanding these differences is critical for financial institutions and model developers, as it highlights the need for region-specific models or the adaptation of existing models to better fit the transaction patterns observed in different markets. This study evaluates the performance of a US-trained multi-class classification model when applied to Indian transaction data. The confusion matrix reveals a significant number of misclassifications, particularly in the "Debt Service" and "Liability" categories. These misclassifications indicate that the model, which was trained on US data, struggles to accurately categorize Indian transactions data. The poor performance in the Indian market can be attributed to several factors, including the differences in keyword distribution and category frequency.



The comparison of keyword frequencies between the US and Indian markets clearly shows that the same set of keywords can have very different distributions depending on the region. This observation has profound implications for the application of machine learning models in different markets. A model trained on US data, for example, may not perform well when applied to Indian data without significant adjustments.

In the study, machine learning models combined with an improved keywords list produced standard results, achieving an accuracy of 83.56% when applied to US data. However, when the same model was stress-tested on Indian data, the accuracy dropped significantly to 66.72%. This decline in performance underscores the challenges of applying a model trained on data from one region to another without considering regional differences in transaction behaviour and keyword usage. The results suggest that for the Indian market, a new set of keywords would need to be developed, tailored specifically to the unique characteristics of Indian transactions. This would require a large amount of data from the Indian market to accurately capture the nuances of keyword usage and transaction categorization. Once these keywords are developed, they could be integrated into the overall keywords master data sheet, allowing for better generalization or the creation of a more generic model that could be applied across multiple regions.

In essence, this study highlights the importance of regional customization in financial modelling. While machine learning models can achieve high accuracy in markets where they have been trained and optimized, their performance can deteriorate when applied to unfamiliar markets with different transaction behaviours. Therefore, to successfully apply these models globally, it is crucial to account for regional differences, whether through the development of new keyword sets, region-specific training data, or adaptive modelling techniques that can adjust to different transaction environments.

## CHAPTER VI:

### SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

#### 6.1 Summary

The study highlights the significant impact of enhancing keyword lists and applying machine learning model to different region like India Lending Market. At the same time study also explore the possibility of using Large Language Models for the classification of the bank statement transactions based on transactions descriptions. The research demonstrates that expanding the keyword list from 1,603 to 3,000 keywords significantly improved the model's performance, boosting accuracy from 73.62% to 83.56%. This finding stress on the importance of a more extensive and refined keyword list in improving the precision of classification model used in this study for bank statements analysis for SME consumers.

The study also explores the potential of Large Language Models (LLMs) in classification tasks, particularly with smaller datasets. Testing the LLM on 10,000 bank transactions yielded an accuracy of 85.20%, suggesting that LLMs are highly effective in handling classification tasks even without extensive training data. The LLM was able to correctly classify 8,129 transactions on the first attempt. The study further tested the model's performance by reclassifying the 1,871 misclassified transactions, of which 391 were correctly classified on the second attempt. This improvement, attributed to the non-deterministic nature of generative AI, demonstrates the LLM's capacity for bank statements transactions classification tasks.

One of the critical insights from the study is the consistency of post-classification processes across different models, including both traditional machine learning and LLM models. This consistency indicates that downstream processes are stable and can handle outputs from various classification models without needing significant modifications.

However, the study also reveals the challenges of applying a model trained in one region to another with different transaction behaviours. When the US-trained classification model was applied to the Indian lending market, its performance dropped significantly, with accuracy falling to 66.72%. This drop illustrates that the keyword list and model optimized for the US market may not be fully applicable to the Indian context due to regional differences in transaction patterns and keywords usage.

The research delves into the process of extracting and enhancing keywords for transaction analysis, particularly focusing on unigrams, bigrams, and trigrams derived from transaction descriptions. The study found that simply selecting the most frequent tokens as keywords was insufficient due to the complex nature of financial terms, abbreviations, and banking terminology present in the data. To address this limitation, domain expertise was used to refine the generated keywords, resulting in a more effective classification model with the enhanced keyword list.

The study also compares the frequency of common transaction keywords in the US and Indian markets, revealing distinct sets of keywords with varying frequencies across the two regions. This difference in keyword distribution highlights the need for region-specific models or the adaptation of existing models to better fit the unique transaction patterns observed in different markets.

Additionally, the study examines the distribution of transactions across five cash flow categories ("Income," "Expense," "Transfers," "Debt Service," and "Liability") in both the US and Indian markets. The US market showed a more balanced distribution across these categories, while the Indian market exhibited a skewed distribution, with a higher percentage of transactions in the "Transfers" and "Liability" categories. These differences suggest that regional banking practices and consumer behaviours vary significantly, further emphasizing the necessity for customized models.

The confusion matrix analysis of the US-trained model's performance on Indian data revealed significant misclassifications, particularly in the "Debt Service" and "Liability" categories. This finding indicates that the model struggles to accurately categorize transactions when applied to a market with different transaction behaviours.

## **6.2 Implications**

This study emphasizes the need for regional customization in order to use bank statement analyzer for different regions. Although machine learning models can attain high accuracy in the markets they are trained in, their effectiveness may drop when applied to different regions without considering local variations in transaction behavior and keyword usage. For example, a model trained on U.S. banking data may underperform when applied to the Indian lending market due to distinct patterns in transactions and keywords. To improve the global applicability of such models, it is essential to develop region-specific keyword sets and incorporate localized training data. Additionally, adaptive modeling techniques can help machine learning systems adjust to diverse financial environments. This approach ensures that models remain accurate and reliable across various regions, reducing the risk of misclassifications in unfamiliar markets. Overall, the study highlights the critical role of regional differences in financial data and emphasizes the importance of customizing models for each market to achieve optimal results.

The study emphasizes the crucial role of expanding keyword lists and adapting machine learning models to new regions, such as the Indian lending market. It investigates the potential of Large Language Models (LLMs) for classifying bank statement transactions based on transaction descriptions. This finding suggests that LLMs are particularly suited for financial transaction classification tasks, given their ability to learn and adapt quickly. The study highlights the importance of combining a comprehensive keyword list with advanced modeling techniques like LLMs for improved accuracy in classifying bank transactions.

### **6.3 Recommendations for Future Research**

The study provides valuable insights into the potential of machine learning and Large Language Models (LLMs) for classifying bank statement transactions, especially with the application to different regions like the Indian lending market. Several key areas emerge from the findings that require further research to improve model accuracy, adaptability, and scalability across regions.

One of the critical findings is the significant drop in model accuracy when a U.S.-trained model was applied to the Indian lending market. The 66.72% accuracy in India, compared to the U.S. performance, highlights the importance of regional customization. Future research should focus on creating region-specific models to better capture the unique transaction patterns, financial terminology, and consumer behaviours in various regions.

As the study demonstrates, keyword selection plays a critical role in model performance. Further research should focus on refining keyword lists for specific regions, possibly through the use of local bank statements keywords extraction. This keywords list should be updated with the keywords for different regions and should be updated using self-learning models. If model comes across any new keyword that keyword should be added to the master keyword list. In this way, more generic model could be attained that could be applied to multiple geographies at the same time with minimal customizations.

The improvement in accuracy, from 73.62% to 83.56%, through an expanded keyword list highlights the importance of more refined keywords. However, the study also reveals the challenge of manually selecting and refining keywords, as domain expertise is needed to capture the nuances of financial terms. Automating this process could significantly enhance both the scalability and efficiency of keyword refinement as explained in above paragraph by using self-learning AI models. Future work should leverage advanced NLP techniques to automatically extract relevant keywords from large datasets of transaction descriptions. Algorithms that consider context and domain-specific knowledge could produce more

refined and accurate keyword lists, improving classification models without heavy reliance on human intervention.

The study reveals that LLMs can achieve high accuracy (85.20%) with smaller datasets (10,000 transactions) and improve after reclassifying misclassified transactions. However, LLMs are resource-intensive and might face limitations when scaling up to larger datasets or regions with different transaction behaviours. Future research should investigate how to maximize the potential of LLMs in bank statement classification tasks. Although LLMs perform well with minimal training data, fine-tuning them on domain-specific datasets could further improve accuracy. Future studies could fine-tune LLMs on large datasets of financial transactions from diverse regions, enabling them to better understand and classify complex bank statements transactions. The study shows that LLMs improve accuracy when reclassifying misclassified transactions. This suggests that a multi-turn classification approach, where LLMs iteratively refine their classifications, could enhance overall performance.

This study highlights significant differences in transaction patterns across regions, such as the skewed distribution in the “Transfers” and “Liability” categories in the Indian market. Future work should expand this comparative analysis to other regions to better understand the global applicability of classification models. Comparative studies involving regions like Europe could provide insights into universal transaction patterns and region-specific outliers. By analysing data from various banking systems, researchers can identify common challenges in transaction classification and propose solutions that apply across different markets. Future research could focus on improving the model to handle misclassifications more effectively, particularly in regions with complex financial behaviours. One way is to explore the development of automated feedback systems where misclassified transactions are flagged for human review or reclassified automatically based on model confidence levels. Such feedback loops can improve model accuracy.

## 6.4 Conclusion

This study emphasized on the vital role that keywords pertaining to the banking transactions plays in the development and application of machine learning models for the classification of the transactions into different categories, especially when dealing with diverse regional markets. One of the key takeaways from this research is the importance of refining and expanding the keyword list used in the classification of banking transactions. By enhancing the list of keywords, the model's performance showed a significant improvement. This improvement, though substantial, is not sufficient to ensure the generalization of the model developed in this study, particularly when applied to markets with distinct transactional behaviors, such as the Indian market compared to the US which was the primary region for which the model was developed.

The research included the enhanced set of keywords used to classify banking transactions from a smaller, less diverse list to a more comprehensive one. This enhancement allowed the classification model to capture a broader range of banking transactions, improving its ability to classify various banking transactions accurately. Input data was having abbreviations, and financial terms, because of this complexity the development of classification model was not straight forward. Moreover the input data was not labelled. And labels are to be determined by using the clustering mechanism.

The second major finding from the study is the performance of Large Language Models (LLMs) in this context. While traditional machine learning models showed marked improvement with keyword list enhancements, LLMs proved to be even more effective in classifying transactions with minimal training. The generative nature of LLMs and their capacity to handle complex, unstructured text provide a significant advantage when working with such complex transactions data, which often lacks uniformity. The high accuracy achieved by LLMs on a relatively small dataset shows their potential as powerful tools in the banking transaction classification, particularly when time and resources for model training are limited.

The process of resubmitting incorrectly classified transactions to the LLM for a second attempt further highlights the adaptability of these models. The non-deterministic nature of LLMs, which allows them to generate different outputs from the same input, enabled the model to classify additional transactions correctly on the second pass. This iterative approach reflects the flexibility of LLMs, which can continuously refine their output, a trait that traditional models typically lack. The high level of accuracy achieved through few-shot learning—where the model improves its performance with minimal additional input—suggests that LLMs may be particularly well-suited for scenarios where the complex unlabeled data is present as in our case in our study.

However, the study also reveals some of the limitations and challenges associated with applying machine learning models across different markets, particularly when it comes to regional differences in bank statements. When the US-based model was applied to the Indian lending market, performance dropped significantly. This underlines the fact that financial transaction patterns and consumer behaviors are not universal. The keywords that are prevalent and meaningful in one market may have little relevance in another. For instance, certain transaction types that are common in the US market may be rare or even absent in the Indian market, and vice versa. This disparity in keyword distribution results in a lower classification accuracy when the model is applied outside its original training environment. The misclassification of Indian transactions when using a model trained on US data illustrates the importance of regional customization. For example, the Indian market demonstrated a skewed distribution of transactions towards certain categories like "Transfers" and "Liability," a pattern not seen in the US data. This suggests that to create a model capable of accurately classifying transactions in the Indian market, or any other non-US market, it would be necessary to develop a region-specific keyword list and model parameters that reflect the unique characteristics of that market.

In essence, this study demonstrates that while machine learning models, including both traditional models and LLMs, did well for the classification of financial transactions, their



success is heavily dependent on the context in which they are applied. In a globalized scenario, where banking transactions can vary widely across regions, a one-size-fits-all model is unlikely to succeed. Instead, models must be tailored to the specific financial terms, transaction types, and consumer behaviors of the regions in which they are supposed to work. This could involve developing distinct models for each region or, alternatively, creating adaptive models that can learn and adjust to new transactional environments over time.

For the Indian market, the study suggests that a more customized approach is needed, one that takes into account the specific transactional behaviors and keyword distributions of the region. Developing a comprehensive, region-specific set of keywords would likely lead to a significant improvement in the model's performance. This effort, however, requires access to a large and diverse dataset that accurately reflects the range of transactions occurring in the market. Once a suitable keyword list is created, it could be incorporated into the model's training process, allowing for better generalization and performance across different regions.

The findings of this study in short elaborates the importance of regional customization, the power of LLMs in handling banking data, and the continued need for enhancing keyword extraction methods in transaction classification. As the financial industry becomes increasingly globalized, the ability to develop and generalize the machine learning models to work for different local markets will be critical. This study serves as a valuable research, illustrating both the opportunities and the challenges of applying machine learning both traditional and advanced (LLM) in the complex scenario of the banking transactions.

## REFERENCES

- Ajithkumar, A. and Geetha, S. (n.d.). “Classification of E-commerce Financial Transaction Logs Using Machine Learning Approach”, *International Journal of Health Sciences*, ScienceScholar, No. I, pp. 4500–4506.
- Rao, P., Kumar, S., Chavan, M. and Lim, W.M., 2023. A systematic literature review on SME financing: Trends and future directions. *Journal of Small Business Management*, 61(3), pp.1247-1277.
- RAJU, K. and REDDY, Y.D., 2023. An Analysis of Financial statements of State Bank of India from 2014 to 2019-----. *The journal of contemporary issues in business and government*, 29(1), pp.497-519.
- Cornelli, G., Frost, J., Gambacorta, L., Rau, P.R., Wardrop, R. and Ziegler, T., 2023. Fintech and big tech credit: Drivers of the growth of digital lending. *Journal of Banking & Finance*, 148, p.106742.
- Mytnyk, B., Tkachyk, O., Shakhovska, N., Fedushko, S. and Syerov, Y., 2023. Application of artificial intelligence for fraudulent banking operations recognition. *Big Data and Cognitive Computing*, 7(2), p.93.
- Goel, A. and Rastogi, S., 2023. Lending to SMEs: Indian banker's perspective. *International Journal of Business and Globalisation*, 34(4), pp.498-514.
- Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T. and Wang, G., 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.
- Beaumont, P., Tang, H. and Vansteenbergh, E. (2022), “The role of fintech in small business lending”, *Available at SSRN 4260842*.

Calabrese, R., Degl'Innocenti, M. and Zhou, S. (2022), "Expectations of access to debt finance for SMEs in times of uncertainty", *Journal of Small Business Management*, Taylor & Francis, Vol. 60 No. 6, pp. 1351–1378.

Chen, B., Zhang, C. and Saydaliev, H.B. (2022), "Does bank complexity during the COVID-19 crisis alter the financing mechanism for small and medium-sized enterprises?", *Economic Analysis and Policy*, Elsevier, Vol. 75, pp. 705–715.

Cornelli, G., Frost, J., Gambacorta, L., Rau, R., Wardrop, R. and Ziegler, T. (2021), "Fintech and big tech credit: What explains the rise of digital lending?", Vol. 22, presented at the CESifo Forum, München: ifo Institut-Leibniz-Institut für Wirtschaftsforschung an der ..., pp. 30–34.

Albarrak, M.S. and Alokley, S.A. (2021), "FinTech: Ecosystem, Opportunities and Challenges in Saudi Arabia", *Journal of Risk and Financial Management*, MDPI, Vol. 14 No. 10, p. 460.

Patel, R., & Gupta, S. (2021). Digital Transformation in SME Lending. *International Journal of Business and Finance Research*, 15(2), 45-622

Anne, J., Gichuki, W., Njeru, D.A. and Tirimba, O.I. (2014), "Challenges Facing Micro and Small Enterprises in Accessing Credit Facilities in Kangemi Harambee Market in Nairobi City County, Kenya", *International Journal of Scientific and Research Publications*, Vol. 4 No. 1, pp. 2250–3153.

Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T.F., Klein, J. and Goujon, A. (2021), "A comparison of pre-trained language models for multi-class text

classification in the financial domain”, presented at the Companion Proceedings of the Web Conference 2021, pp. 260–268.

Baharudin, B., Lee, L.H. and Khan, K. (2010), “A Review of Machine Learning Algorithms for Text-Documents Classification”, *Journal of Advances in Information Technology*, Vol. 1 No. 1, available at:<https://doi.org/10.4304/jait.1.1.4-20>.

D. Ghosh, A. Gupta, K. gupta. (2019), “Unlocking Untapped Potential”, *NASSCOM*, No. August.

Dr.S.Poongavanam. (2017), “An assignment on A STUDY ON CASH FLOW STATEMENT ANALYSIS WITH SPECIAL REFERENCE TO JET AIRWAYS”, *IOSR Journal of Business and Management (IOSR-JBM)*, Vol. 22 No. 10, pp. 127–143.

Egesa, A. (2016), “Analysis of Financial Transactions using Machine Learning”, *Lu-Cs-Ex 2016-05*.

E&Y. (2018), “The future of SME banking”, No. December.

Fiengenbaum, A. and Thomas, H. (1986), “An empirical study”, *Strategic Management Journal*, Vol. 7 No. 1, pp. 395–407.

FinTech, F. (2019), “SME lending: banks aren’t using enough data – FinTech Futures””, available at: <https://www.fintechfutures.com/2019/08/banks-arent-using-enough-data/>.

Frame, W.S., Srinivasan, A. and Woosley, L. (2001), “The effect of credit scoring on small-business lending”, *Journal of Money, Credit and Banking*, JSTOR, pp. 813–825.

Futures FinTech. (2019), “SME lending: banks aren’t using enough data – FinTech Futures”, *Www.Fintechfutures.Com*, available at: <https://www.fintechfutures.com/2019/08/banks-arent-using-enough-data/> (accessed 9 February 2022).

Gopal, M. and Schnabl, P. (2022), “The rise of finance companies and fintech lenders in small business lending”, *The Review of Financial Studies*, Oxford Academic, Vol. 35 No. 11, pp. 4859–4901.

Guidotti, R., Monreale, A., Nanni, M., Giannotti, F. and Pedreschi, D. (2017), “Clustering individual transactional data for masses of users”, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. Part F1296 No. August, pp. 195–204.

Haralayya, B. (2021), “How Digital Banking has brought innovative products and services to India”, *Journal of Advanced Research in Quality Control and Management*, Vol. 6 No. 1, pp. 16–18.

Huang, J., Chai, J. and Cho, S. (2020), “Deep learning in finance and banking: A literature review and classification”, *Frontiers of Business Research in China*, SpringerOpen, Vol. 14 No. 1, pp. 1–24.

Jørgensen, R.K. and Igel, C. (2021), “Machine learning for financial transaction classification across companies using character-level word embeddings of text fields”, *Intelligent Systems in Accounting, Finance and Management*, Wiley Online Library, Vol. 28 No. 3, pp. 159–172.

K. Kahnamouei, T. (2017), “Providing a Model to Predict Future Cash Flow Using Neural Networks on the Pharmaceutical and Chemical Industries of Tehran Stock Market”, *International Journal of Accounting and Financial Reporting*, Vol. 7 No. 1, p. 213.

Kowsari, K., Meimandi, K.J., Heidarysafa, M., Mendu, S., Barnes, L. and Brown, D. (2019), “Text classification algorithms: A survey”, *Information (Switzerland)*, Vol. 10 No. 4, pp. 1–68.

Kumar, B.S. and Ravi, V. (2016), “A survey of the applications of text mining in financial domain”, *Knowledge-Based Systems*, Elsevier, Vol. 114, pp. 128–147.

Leo, M., Sharma, S. and Maddulety, K. (2019), “Machine learning in banking risk management: A literature review”, *Risks*, MDPI, Vol. 7 No. 1, p. 29.

Lokmic, L. and Smith, K.A. (2000), “Cash flow forecasting using supervised and unsupervised neural networks”, *Proceedings of the International Joint Conference on Neural Networks*, Vol. 6 No. February 2000, pp. 343–347.

Mahalakshmi, B. and Duraiswamy, K. (2012), “An Overview of Categorization Techniques”, *International Journal of Modern Engineering Research (IJMER)*, Vol. 2 No. 5, pp. 3131–3137.

Merz, M. and Riepe, J. (2021), “SMEs with legally restricted banking access: evidence from the US marijuana industry”, *Journal of Business Economics*, Springer, Vol. 91 No. 6, pp. 797–849.

Mills, K.G. (2018), *Fintech, Small Business & the American Dream: How Technology Is Transforming Lending and Shaping a New Era of Small Business Opportunity*, Springer.

Nassoro, G. and Jaraj, K. (2022), “Challenges small and medium enterprises (SMEs) face in acquiring loans from commercial banks in Tanzania”, *African Journal of Business Management*, Academic Journals, Vol. 16 No. 4, pp. 74–81.

Omran, M.G.H., Engelbrecht, A.P. and Salman, A. (2007), “An overview of clustering methods”, *Intelligent Data Analysis*, Vol. 11 No. 6, pp. 583–605.

P. Bhakta. (2019), “Digital lending startup: Here’s why many startups are going into the ‘risky’ business of lending - The Economic Times”, *Economictimes*, available at: <https://economictimes.indiatimes.com/small-biz/startups/features/heres-why-many-startups-are-going-into-the-risky-business-of-lending/articleshow/68214436.cms?from=mdr> (accessed 9 February 2022).

Rao, P., Kumar, S., Chavan, M. and Lim, W.M. (2021), “A systematic literature review on SME financing: Trends and future directions”, *Journal of Small Business Management*, Taylor & Francis, pp. 1–31.

Saal, M., Starnes, S. and Rehmann, T. (2017), “Digital Financial Services”, International Finance Corporation, Washington, DC.

Semenyuta, O., Andreeva, A.V., Sichev, R. and Filippov, Y.M. (2019), “Digital technologies in lending small and medium-size enterprises in Russia”, Eleftherios Thalassinou.

Anon (2020) *DBSCAN - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/DBSCAN> [Accessed 16 Feb. 2023].

Heidenreich, H., (2020) *Natural Language Processing: Count Vectorization with scikit-learn*. [online] towardsdatascience.com. Available at: <https://towardsdatascience.com/natural-language-processing-count-vectorization-with-scikit-learn-e7804269bb5e> [Accessed 16 Feb. 2023].

Maklin, C., (2019) *TF IDF | TFIDF Python Example - Towards Data Science*. [online] towardsdatascience.com. Available at: <https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76> [Accessed 16 Feb. 2023].

Srinidhi, S., (2019) *Understanding Word N-grams and N-gram Probability in Natural Language Processing*. [online] towardsdatascience.com. Available at: <https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058> [Accessed 16 Feb. 2023].

Yiu, T., (2019) *Understanding Random Forest - Towards Data Science*. [online] towardsdatascience.com. Available at: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2> [Accessed 16 Feb. 2023]

Singh, J. (2018), “Challenges in Text Classification Using Machine Learning Techniques”, *International Journal of Recent Trends in Engineering and Research*, Vol. 4 No. 2, pp. 81–83.

Mehta (2020), “Bank Statement Analyser”, No. February.

Sun, T. (2021), “Digital Banking Support to Small Businesses amid COVID-19”, *Global Financial Stability Notes*, International Monetary Fund, Vol. 2021 No. 002.

Thangaraj, M. (2018), “Text Classification Techniques : a Literature Review”, Vol. 13, pp. 117–135.



Vollset, E.E.N. and Folkestad, O.E.E. (2017), “Automatic classification of bank transactions”, NTNU.

Wadesango, N., Tinarwo, N., Sitcha, L. and Machingambi, S. (2019), “The impact of cash flow management on the profitability and sustainability of small to medium sized enterprises”, *International Journal of Entrepreneurship*, Vol. 23 No. 3.

Yadav, S. and Thakur, S. (2018), “Bank loan analysis using customer usage data: A big data approach using Hadoop”, *2nd International Conference on Telecommunication and Networks, TEL-NET 2017*, Vol. 2018-Janua, pp. 1–8.

APPENDIX A:  
ENHANCED KEYWORDS LIST

Enhanced keywords file is placed below for reference:



EnhancedKeywordsLi  
stv1.1.xlsx