FINAL CAPESTONE PROJECT BRIEF

Customer Financial Risk Prediction & Sentiment Analysis

Domain: Finance

Techniques: Clustering, Topic Modeling, Sentiment Analysis

Tools: Python, Power BI, NLP (spaCy / Transformers), K-Means, PCA

Dataset: 5,000+ financial behavior records

1. Background

Across African financial markets—such as **Nigeria, Ghana, Kenya, Côte d'Ivoire, Uganda, and South Africa**—institutions struggle to understand customer financial behaviour due to:

- Inconsistent or incomplete financial histories
- Rapid adoption of digital payments (USSD, POS, Mobile Money, wallet apps)
- High levels of informal income and cash-based transactions
- Unstructured customer feedback (calls, SMS, WhatsApp chats)
- Limited use of traditional credit scoring frameworks

As a result, **Banks, Microfinance Institutions, Digital Lenders, and SACCOs** increasingly rely on analytics-driven systems that provide:

- Behaviour-driven segmentation
- Wallet and account activity improvement
- Personalized product recommendations
- Targeted financial education
- Early detection of potentially risky behavioural patterns (even without predefined labels)

This project delivers a **fully unsupervised ML + NLP solution** built specifically around African financial behaviour patterns.

2. Project Objective

To develop a machine learning and NLP system that groups customers based on:

- Monthly expenditure and spending patterns
- Digital payment behaviour (POS, USSD, Transfers, Mobile Money)
- Savings consistency and income stability
- Category-level spending (food, rent, data, transport, utilities)
- Customer sentiment extracted from feedback
- Complaint themes/topics discovered using NLP
- Lifestyle-linked financial behaviour trends

No labels are provided.

Instead, the system uncovers **natural, data-driven customer segments** and identifies hidden behavioural patterns.

3. Dataset Description (Matches Your 5,000+ Dataset)

A. Demographics & Financial Profile

Feature	Description
customer_id	Unique user ID
age	18–70 years
income_level	Low / Middle / High
monthly_expenditure	Total monthly spend (NGN/KES/GHC/ZAR equivalent)
saving_behavior	Consistent / Irregular / None
credit_score	300–850 (semi-structured credit proxy)

B. Digital Payment Behaviour

Feature	Description
transaction_count	No. of transactions per month
avg_transaction_value	Average value per transaction
payment_channels_used	POS, Transfer, USSD, Mobile Money, ATM withdrawals
expenditure_categories	Food, Data, Utilities, Rent, Fuel, Transport, etc.

C. NLP Data

Feature	Description
customer_feedback	Raw review or complaint text
sentiment_score	NLP-derived score (–1 negative to +1 positive)
topic (derived later)	Extracted using LDA or BERTopic

4. Methodology (End-to-End Workflow)

Step 1 — Data Cleaning

- Handle missing income/savings values
- Normalize monthly expenditure
- One-hot encode payment channel usage
- Clean NLP text (tokenization, stopwords, punctuation removal)
- Standardize all numeric features

Step 2 — NLP Processing

- Lemmatization
- Sentiment analysis (spaCy / VADER / transformers)
- Topic modeling (LDA / BERTopic)
- Generate embeddings (TF-IDF or Sentence Transformers)

Step 3 — Feature Engineering

- PCA on numeric data
- UMAP/PCA on text embeddings
- Merge structured and NLP features into a unified matrix

Step 4 — Unsupervised Modeling

Models to evaluate:

- **K-Means Clustering** (baseline)
- **Hierarchical Clustering**
- **Gaussian Mixture Models**
- **DBSCAN** (for detecting behavioural anomalies/outliers)

Step 5 — Cluster Validation & Interpretation

Clusters may naturally represent:

- **High Spenders with Positive Sentiment**
- **Low Income, Irregular Savings, High Complaints**
- **Digital-First, Mobile Money Heavy Users**
- **Stable Earners with Consistent Savings**
- **Cash-Based, Low Digital Adoption Users**

Cluster names are assigned after analysis of cluster features.

6. Power BI Dashboard (Analytics Team Responsibility)

The analytics team will design a **Power BI dashboard** that displays:

➤ **Customer Segmentation**

- Size of each cluster
- Behavioural summaries per segment

➤ **Payment Channel Analytics**

- POS vs Transfer vs USSD usage
- Digital adoption trends

➤ **Financial Behaviour Metrics**

- Average monthly expenditure per cluster
- Savings consistency by cluster
- Credit score distribution

➤ **NLP Insights**

- Sentiment trend across clusters
- Top complaint topics
- Emerging customer issues

➤ **Cluster Evolution Over Time**

- Movement between clusters

- Appearance of new behaviour segments
- Early signals of potentially high-risk patterns

7. Deliverables

✓ 1. Power BI / Streamlit Dashboard

Interactive dashboard showing clusters, sentiment trends, and financial patterns.

✓ 2. Model Deployment (FastAPI)

A deployed unsupervised + NLP inference API that:

- Accepts new customer data
- Assigns cluster membership
- Computes sentiment & topic
- Returns behavioural insights

✓ 3. Presentation Slide Deck

A professional summary covering:

- Problem statement
- Dataset overview
- Methodology
- ML/NLP results
- Dashboard demo
- Recommendations for the financial institution