



ASHESI UNIVERSITY

**MOBILE MONEY SMS CLASSIFICATION AND TEXT
ANALYSIS: EXPLORING POSSIBILITIES FOR ENHANCED
FINANCIAL INCLUSION**

Applied Project

BSc. Computer Science

Sihle Magagula

2020

Ashesi University

**Mobile Money Sms Classification and Text Analysis: Exploring
Possibilities For Enhanced Financial Inclusion**

APPLIED PROJECT

Applied Project submitted to the Department of Computer Science, Ashesi University in
partial fulfilment of the requirements for the award of Bachelor of Science degree in
Computer Science

Sihle Magagula

May 2020

DECLARATION

I hereby declare that this applied project is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature:

.....

Candidate's Name:

.....

Date:

.....

I hereby declare that preparation and presentation of this applied project were supervised in accordance with the guidelines on supervision of applied project laid down by Ashesi University.

Supervisor's Signature:

.....

Supervisor's Name:

.....

Date:

.....

Acknowledgements

This applied project would not have been possible without the selfless contributions of a few of people. First and foremost, I would like to thank my supervisor, who's academic advice and rigor helped me undertake this project. I would also like to thank the friend who generously offered his SMSs to be used for this applied project. Finally, I thank my family and friends and a few individuals in the Ashesi community for their boundless support.

Abstract

In the past two decades, financial technology (fintech) has grown to become one of the most significant economic drivers in developing countries especially in sub-Saharan Africa. Despite the prevalence of these fintech mostly in the form of mobile money platforms, the number of unbanked populations across developing countries has remained high. This applied project presents a human-centered approach in the innovator-side exploration of the integration between the banking sector and fintech. Such innovations should ask nothing more from the user than they already have, should adopt a fluid digital footprint, and the services offered by integrated platforms should be dynamic. To that end, the paper presents a system that classifies mobile money SMSs and use them to prepare a secure financial statement that might enhance the Know-Your-Customer requirements for the unbanked and also ensure that they can easily transfer their mobile money credit record and easily access services in the banking sector.

Keywords:

Text Classification; text analysis; machine learning

Table of Contents

DECLARATION	iii
Acknowledgements	iv
Abstract	v
Table of Contents	vi
Chapter 1: Introduction	1
1.1 Background: Mobile Money in Eswatini.....	4
1.2 Related Work.....	5
1.2.1 Mobile Money Fraud.....	5
1.2.2 Mobile Money Adoption and Expansion.....	5
1.2.3 Financial Inclusion via Mobile Money.....	6
1.2.4 Integration: M-Shwari and M-Pawa.....	7
Chapter 2: Requirements and Specifications	8
Overview.....	8
2.1.1 Objective.....	8
2.1.2 Approach.....	8
2.2 Requirement gathering procedures.....	9
2.3 User classes and Use cases.....	10
2.4 Functional requirements.....	10
2.4.1 Large dataset.....	10
2.4.2 High accuracy.....	10
2.4.3 Synthesizing a presentable document format.....	11
2.4.4 Security.....	11
2.4.5 Verification.....	11
2.5 Non-functional requirements.....	11

2.5.1 Reliability	11
2.5.2 User friendly	12
Chapter 3: Architecture and Design	13
3.1 System Overview	13
3.1.1 High Level Architecture	13
3.2 Key Components	13
3.2.1 Data preparation	14
3.2.2 Model Training	14
3.2.3 Mobile System.	15
Chapter 4: Implementation	16
4.1 Datasets	16
4.1.1 Data Mining	16
4.1.2 Format and Labelling	17
4.1.3 Data Augmentation	18
4.1.4 Data Occlusion.	20
4.2 Classification Model	20
4.3 Text Analysis.	21
4.3.1 Formatting and Cleaning	21
4.3.2 Analysis	21
Chapter 5: Testing and Results.	25
5.1 Text Classification.	25
5.2 Text Analysis	28
Chapter 6: Conclusions and Future Work.	30
References	31

Chapter 1: Introduction

In the past two decades, financial technology (fintech) has grown to become one of the most significant economic drivers in developing countries especially in sub-Saharan Africa. The most noticeable of these technologies has been the almost ubiquitous mobile money platforms offered by Mobile Network Operators (MNOs) like MTN and Vodafone. To date, there are more than 395.7 million registered mobile money accounts in sub-Saharan Africa which represents more than half of the total globally [11]. On average, these represent more than 60% of the adult population in most countries. It is projected that the region will have more 600 million unique subscribers to mobile services like mobile money in 2025.

Despite the prevalence and sustained growth of mobile money platforms, the number of unbanked populations across developing countries has remained high. According to the World Bank there are still a billion people in the world who are unbanked. Furthermore, there are still 1.7 billion people around the globe who do not have access to safe, reliable, and convenient financial services [7]. This means that almost one fifth of the global population is financially excluded.

With the already ubiquitous fintech like mobile money platforms, is there more that can be done to catapult these populations closer to a financial safety net and ensuring financial inclusion for all? There are promising solutions to this question in some of the countries that currently lead innovation in this space like Kenya and Ghana. For instance, in Kenya there's M-Shwari a mobile application that is connects to a user's mobile money account and a bank account and allows the user to have access to quick loans from the bank [5]. This is a great case of how the populace can be financially included through some of the available fintech.

Unfortunately, there remain several obstacles for innovation and adoption in this arena. The first and probably obvious one is that like all other societal challenges there is no one-size-

fits-all solution for financial inclusion. A significant share of this problem is market oriented. Promising solutions like M-Shwari cannot achieve meaningful progress alone. There needs to be many solutions targeted at the same problems and the invisible hand of the market will guide users to the best solutions that meet their needs. For instance, M-Shwari is quick loan service, but what about users who need insurance or mortgage solutions? This space is in need of many more solutions and fortunately several companies have thrown their weight to the problem by sponsoring innovation through fintech challenges [12].

Related to the market problem and actually hinged to it is the policy problem in this particular innovation space. There's no doubt that policymakers will play a big part if the financial exclusion is to be conquered. Their decisions matter both for the innovators and the consumer. The World Bank has taken the first step in stimulating governments into action on this part. Its formulation of the Universal Financial Access 2020 initiative has prompted some sub-Saharan governments into action. For example, the Eswatini Government met the World Bank's initiative by a national strategy to combat financial inclusion though crisis like Covid-19 pandemic might reverse the gains on this front [8].

The last obstacle, which is also the focus of this paper, is the integration of fintech and traditional banking institutions has been certain but slow. The speed of technological innovation is usually fast and its truer for fintech. But what remains the challenge at the moment is the rigidity and high regulation standards of the banking sector [4]. This is a huge barrier of entry for independent innovators. No wonder thriving solutions like M-Shwari are an inhouse innovation through a collaboration of an operator of mobile money, M-Pesa and a bank, Commercial Bank of Africa [5]. On the contrary, this instance reveals that the integration of the banking sector and fintech is possible and the possibilities would be limitless if the pace of integration would accelerate.

This paper presents an innovator-side exploration of the integration between the banking sector and fintech. The presented solution makes use of machine learning to classify Mobile Money SMSs generated on the user-end of the mobile money platform from other SMSs that a user might receive. The purpose of this classification is to explore the useful data in the SMSs and prepare a financial statement that might be presented to lenders like those in the banking sector.

The potential solution presented here takes an approach that, we contend, should guide all solutions to the challenge of integrating the banking sector and fintech. The approach is consumer oriented in the following three ways. First, it asks nothing more of the user other than consent to classify their SMSs on their personal device. The assumption is that the user already uses a mobile money platform and if that's true they have a mobile phone. They also have the SMSs that are rarely used after a transaction on the mobile money platform has been confirmed and completed.

Second, the approach is fluid. Too often a digital footprint of consumer is stuck in one platform or ecosystem and cannot be transferred to other systems where it might actually serve the consumer. With our potential solution we make use of the digital footprint created in mobile money platforms to create access to opportunities in the banking sector. Usually upon opening a bank account, users are made to wait until they have a sufficient track record or provide further documentation to have access to higher order services like loans and mortgage. But that need not be the case with the tentative solution presented in this paper. They can transfer the record they already have with the mobile money platform and access any services they qualify for right away.

The third and last aspect of our approach has the potential to be dynamic. The transfer of the digital footprint from one platform to the end should not be the end. The transfer of digital footprint should always be ready and done with the consent of the consumer. As the

consumer continues to make use of one platform, they should be capable of using that updated digital footprint to inform the other of the new information that might lead to better services.

The study is a case of Eswatini. The target is the unbanked population that actively transacts via mobile money platforms like that which is provided by MTN Eswatini. This applied project provides the following contribution: explores and presents a potential approach to the banking sector and fintech integration, presents a system that uses machine learning to classify mobile money SMSs from other SMSs and then analyze these SMSs to produce a financial statement.

1.1 Background: Mobile Money in Eswatini

Mobile Money made its first market entry in Eswatini (formerly known as Swaziland) in 2011 back when MTN was the sole mobile network operator (MNO). Now there are two MNOs but MTN overwhelmingly dominates the mobile money market. The MTN mobile money platform has more than 600,000 subscribers which is almost the entire adult population. More than 400,000 of these subscribers are active. At first, subscribers were only allowed to transact a maximum amount of US \$250 per day but that amount was revised in 2018 and now subscribers can transact up to US \$1370 [citation].

While the endeavor of getting all adults subscribed to the MTN's mobile money platform has been an overwhelming success, the same cannot be said about the integration of the platform with the financial sector. The first steps towards integration came in 2017 when MTN partnered with a local bank, Swaziland Building Society (SBS), to allow account holders of SBS and MTN to transfer money between these accounts [9]. Furthermore, SBS account holders could withdraw from their mobile money account via an ATM. Partnerships with other local banks have followed suit with some limited to just withdrawals via ATMs without the ability to transfer funds between the bank account and the mobile money account. The

challenge that has remained is the absence of fluidity of the traditional services of one platform to the other. Also, the financially excluded and unbanked populations cannot take advantage of these limited innovations.

1.2 Related Work

1.2.1 Mobile Money Fraud

One of the greatest concerns when it comes to transfer of money in any platform is and has always been security or rather the guard against fraud. For this reason, there is a growing plethora of studies on fraud when it comes to the mobile money industry. On the one hand you have studies that are helping guard against and detect fraudulent mobile money transfers at the platform level [3]. On the other hand, there are studies looking to control fraud at a policy level seeking to build capacity against fraud from the mobile agents to the companies that provide these services by ensuring secure algorithms [1]. At the policy level, there's little regard for data that is generated by these mobile money platforms, instead the focus is building an ecosystem that is resilient to fraud. At the platform level, the data generated is key. For example, [3] proposed a pattern recognition model with the aim of predicting fraudulent mobile money transactions. To build this model, synthetic data was used due to the sensitive nature and inaccessibility of the original transaction SMSs. The question, therefore, is can we do more especially in the direction of empowering the users of such platforms.

1.2.2 Mobile Money Adoption and Expansion

Due to the high adoption and success of one mobile money service by Vodafone in partnership with Safaricom called M-Pesa in Kenya there's been an obsession about expanding mobile money services to the poor especially the unbanked. This is because mobile money

services are highly accessible when compared to formal financial institutions and this is truer in developing countries. While seeking to expand the adoption of M-Pesa, Vodafone collaborated with researchers to perform a quantitative analysis of some data from their M-Pesa product. This data included anonymized phone calls and M-Pesa transactions and machine learning models were used to predict the adoption of M-Pesa and mobile money spending. All this in the effort to design better mobile money systems in developing countries. The challenge with such studies is that they are not easily verifiable and extendable. The datasets used aren't made publicly available. Also, they tend to be oriented towards the company goals rather than the true needs of the users. In this applied project, the goal is to build models that will enable the poor to access a wider range of financial products via formal financial institutions. To eradicate the market failures of mobile money transfers which are lack of insurance, savings, and credit, there is a need for a holistic and systematic coordination of mobile money ecosystem and that means mobile networks working with formal banking institutions and developers to achieve more [4].

1.2.3 Financial Inclusion via Mobile Money

There have been limited services/applications that build on top of mobile money platforms to expand the financial products that are accessible to the poor. Even when there are such applications, they are usually developed inhouse by the same corporations that own the mobile money service since they have access to the datasets needed to develop such products. One of this application is M-Shwari which is owned by Safaricom, the provider of M-Pesa, in collaboration with Commercial Bank of Africa (CBA). M-Shwari allows M-Pesa users to save and borrow money which includes emergency loans. The thorny issue about such services has been a regulatory one [6]. The current financial system does not cater for mobile savings and credit especially on mobile money platforms like M-Pesa and that makes it hard to build

systems for such scenarios. Another issue related to regulations is how the consumer's data is protected, credit reported, and how transactional data is supposed to be used by the corporations that run the mobile money services such as Safaricom and MTN. It seems the policy choices that must be made aren't too compatible. On the one hand, there's the desirable financial inclusion and on the other, there is the sensitive issue of data privacy and consumer protection.

1.2.4 Integration: M-Shwari and M-Pawa

M-Shwari is a bank account that is offered by the Commercial Bank of Africa (CBA) to M-Pesa users in Kenya. The account allows users to manage their bank account via the M-Pesa wallet; deposits and withdrawals can only be done via M-Pesa. These bank accounts are opened on the basis that a user is already an M-Pesa subscriber and as such there's no need to produce Know Your Customer (KYC) documentation as Safaricom provides all the transaction history and its associated credit scoring to CBA. Ironically, the savings balance users can have in their account can be limited by the amount of KYC documentation they submit [5]. For more savings balance users have to verify their National ID or physically submit their national ID to the bank and for unlimited savings balance, users have to submit their tax ID. Perhaps what's more important than the bank requiring further documentation is that M-Shwari allows the unbanked M-Pesa to have a bank account.

A similar product was launched in neighboring Tanzania and offered the same products; loans and savings. Even though such applications do enable financial inclusion they do little to build the financial capability of the users which is critical when users suddenly have more resources in their hands. Furthermore, the mobile money ecosystem in Swaziland is beginning to realize all these benefits that can be built around the mobile money platform and this project is planning on taking advantage of that.

Chapter 2: Requirements and Specifications

2.1 Overview

2.1.1 Objective

The objective of this applied project is to prototype a system with two overarching goals. The first is to help the unbanked access formal banking institutions by using their transaction records of mobile money transfers to create a verifiable document that can be presented as part Know Your Customer, KYC, procedure. The second is to build financial capacity among the unbanked. By analyzing their mobile money transaction history, the users will be made aware of the patterns that emerge from their transactions and can also be advised on how to either reduce their expenditure, save more money, etc.

2.1.2 Approach

To achieve the set goals, there needs to be big data which can be analyzed and due to the highly regularized nature of the necessary data, synthetic data will be used to train the machine learning algorithms. The synthetic data will resemble the SMS texts that mobile money users receive after any transaction.

The first part of the machine learning algorithm will be to **classify** the SMS texts; some may be non-financial, some financial and some might be spam. The goal here will be to identify the financial SMS texts from the mobile money provider. Once those SMSs are identified the next goal will be to extract the transactional information like the money received or sent and tall it up against expenditures and income and that information will be synthesized to a document with a verification feature like a barcode or any other security feature that will meet the standards that is required by the formal institutions.

The other goal, after classification and related to building financial capability, will be analyze the transactions of each user and make the user aware of the patterns of his transactions and from those patterns the user might make better financial decisions.

2.2 Requirement gathering procedures

The core requirements for this applied project stem from three documents of varying origins but one underlying theme which is understanding the unbanked populations and making every person is financial included. The first is a FinMark's Eswatini FinScope Consumer Survey of 2018 which benchmarks the country's level and quality of financial inclusion and financial capability of consumers. Furthermore, this study gives insights to consumer attitudes and behavior towards the financial sector both formal and informal. The FinMark study is geared towards enabling the improvement of financial services and this project shares that in that goal which is why it draws its requirement from it.

The second document is the World Bank's Universal Financial Access by 2020 which has the aims of making sure that adults without a transaction with a formal financial institution have one by 2020 and are able to receive and send money. Through the World Bank's interventions many partners like MasterCard Global and Equity Bank who are working on the ground to see to it that all citizens of the word have access to a bank account. This project on the other hand seeks to expand bank services to those who have access to mobile money.

The third and final document is the National Financial Inclusion Strategy for Swaziland 2017 – 2020 which lays down the framework at policy level for expanding financial inclusion and financial capability. This document is a product of Universal Financial Access by 2020 where the World Bank partnered with governments to formulate such financial inclusion strategies. It is such strategies that the birth of the Center for Financial Inclusion in Swaziland.

What this document does which is of utmost importance to this project is that it outlines the barriers to financial inclusion some of which regulatory and access related.

2.3 User classes and Use cases

There are two classes of users for this system. The primary class of users are the unbanked people and the secondary class of users are Micro Small Medium Enterprises (MSMEs) who currently have no access to credit. Both classes of users will use the system to access credit with formal banking institutions and also get a get better peek at their transactions.

2.4 Functional requirements

2.4.1 Large dataset

The prerequisite for this for this project as with any other machine learning projects is big data. To be more precise, this project requires a bulk of SMSs both mobile money and random SMSs. The random SMSs are required because at first the algorithm will be taught how to classify between mobile money SMSs and random SMSs. The definition of dig data is not what is clearly defined in the machine learning community but according to some measure it should data sets that are either too large or too complex for traditional data processing systems. The largest SMS corpus available so far is the National University of Singapore SMS Corpus which contains 67,000 SMSs [10]. This means any data set to be used in this project must by some measure attempt to equal or surpass this data set.

2.4.2 High accuracy

While on classification problems the best accuracy score is 100% in practice this score is impossible to achieve. The reasons for this is that the data sample may not be complete, there

might be noise in the data, and the stochastic nature of the modelling algorithm. The target accuracy for this project is anything above 90%.

2.4.3 Synthesizing a presentable document format

One of the goals for this project is to synthesize the correctly classified SMS by extracting expenditure and incomes and creating summary of these in the form of a presentable and readable format like text or pdf.

2.4.4 Security

Though in most cases security is considered as a nonfunctional require in this project it is considered as a functional require for two reasons. First, the nature of the data is highly regularized which means that it contains personal information that might be damaging to individuals if leaked. Secondly, users may attempt to inflate their incomes and expenditures which might not reflect their actual transactions.

2.4.5 Verification mechanism

Related to security is verification. With most documents and any other thing that formal financial institution require must verifiable. And this project the synthesized document should be verifiable to that user who presents it via a certain code.

2.5 Non-functional requirements

2.5.1 Reliability

The system should work well without inconsistencies.

2.5.2 User friendly

The whole system to user friendly to its target group and more so in this case since the target group is expected to be of low literacy.

Chapter 3: Architecture and Design

3.1 System Overview

This system has four main components which are the datasets, trained model, text analysis and mobile application and understanding each component and how it interacts with other components in the system is the goal of this chapter.

3.1.1 High Level Architecture

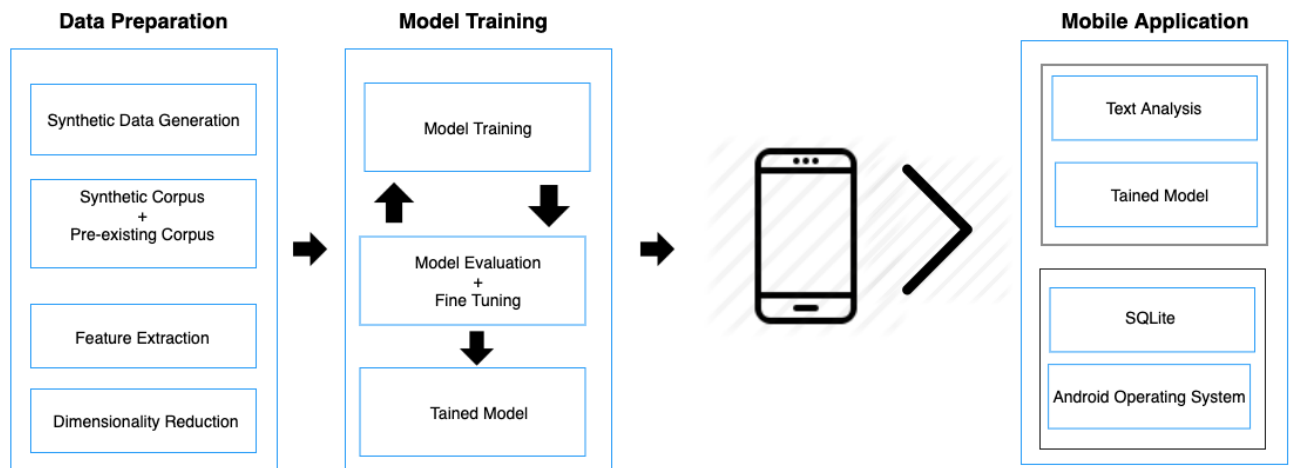


Figure 1: A high level representation of the system

3.2 Key Components

The components approach of this system reduces complexity and perhaps more importantly, takes into account the dependencies among the various functions of the system. For example, the classification of SMSs which is one of the important components has to take

place first before text analysis can be done. Otherwise, if not, that will lead to irrelevant text analysis that will not meet the goal of the system.

3.2.1 Data preparation

Besides preparing the data for training, the data itself has to be generated synthetically. The data that will be generated are the mobile money SMSs. Their deterministic nature makes them a little easy to generate. This involves obtaining a few mobile money SMSs which will be used as templates, anonymizing personal information, and letting figures like dates and amounts float between predetermined ranges.

The second part of data preparation is feature engineering. This involves text cleaning which will remove unnecessary special characters that might be interpreted otherwise by the system. Next, there will be tokenization, a method that breaks down sentences into tokens which are then analyzed either individually or in relation to other tokens. The goal of this form of analysis is to extract key financial information that will be presented in a financial statement.

3.2.2 Model Training

For training the classification model, plain neural networks will be used. Neural networks are good for classification problems and the hope is that they will fare well with SMS classification. The semi-structured nature of the data should boost the performance of the neural network even though it is sequential. Besides SMSs being a form of sequential data, the user will be receiving SMSs over a long period of time and the classifier has to keep track of that.

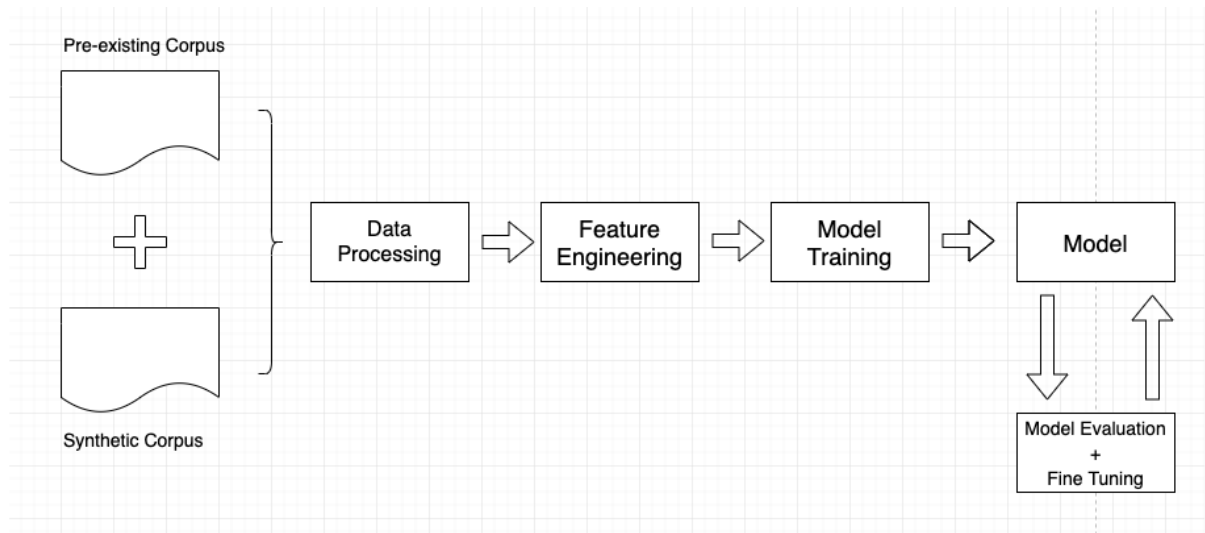


Figure 2: Data processing and model training

3.2.3 Mobile System

Once the model has trained it will be ported to a mobile system where it will work side by side with a text analysis algorithm to meet the goals of this project which are producing a financial statement. What's worth mentioning is that the system will require access to the users' SMSs and storage. The system will access SMSs via SQLite which is the database for android phones.

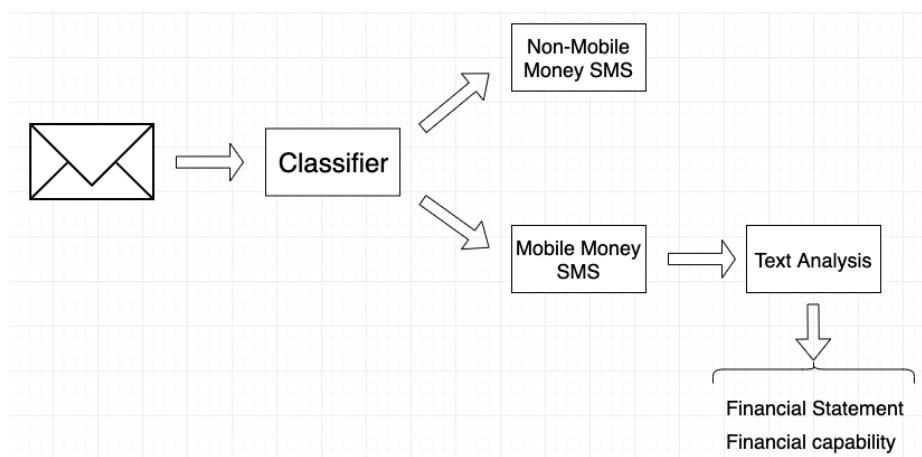


Figure 3: How the model will classify and analyze SMSs in the mobile application

Chapter 4: Implementation

4.1 Datasets

4.1.1 Data Mining

Data was critical to meeting the objectives of this applied project. Even though there has been many open libraries and datasets which have been released by entities for developers to tinker, SMSs datasets remain scarce. This is truer for Mobile Money SMSs because they often times contain sensitive information like names, phones numbers, and account details. This is not to mean that there have been no studies or tinkering in this area of study. Most of the studies almost always withheld their data or are conducted with the holders of this data like MTN and Vodafone both who operate the two largest mobile money platform in Africa.

The available and freely accessible dataset was the University of Singapore SMS Corpus that contained over 60,000 SMSs. Despite its accessibility the data was not suited for this study for two reasons. First it had no mobile money SMSs and it would not be useful to this study. Second, even there were SMSs that might have helped in the generalization of the model, the context of those SMSs has no relevance to the context of this study. Other datasets in platforms like Kaggle also fell short for the same reasons above.

The author then resorted to crowdsourcing. Friends who held a mobile money accounts were identified and approached. They were briefed on the study and asked to contribute their SMSs in confidence that the author would use their SMSs for only this study and also maintain privacy of such data. The following steps outline the steps the select friends had to take for the author to be in possession of such data:

1. Download SMS Backup and Restore, a mobile application in Google Play Store.
2. Use the application to back up their SMSs to their preferred cloud storage service
3. Share the link to that backed up folder of their SMSs

4.1.2 Format and Labelling

The SMS folder were then downloaded and stored locally. The SMSs came in a xml format which is semi-structured data. The next step was to label the data to achieve supervised learning on the classification model. Labelling is one of the demanding tasks in any machine learning pipeline and it was the same for this project. It required that each SMS be copied to an excel spread sheet and be assigned a label. There were two classes of SMSs:

1. MoMo: SMSs that were from the mobile money platform
2. Ham: any other SMS that was not from the mobile money platform

There were eight classes of MoMo SMSs:

1. Bills (electricity, water, etc.)
2. Regular transfer for outgoing and incoming
3. Cash-in or cash-out via an ATM or an agent
4. Loan acquisition or repayment
5. Services like data bundle, airtime, and call minutes purchases
6. Promotional
7. Gifted services
8. Notifications

```

MoMo_v2.tsv x MoMoClasses.xml x
1 Regular
2 <sms protocol="0" address="M-Money" date="1553183698797" type="1" subject="null" body="You have received 1300.00 SZL
from ***** on your mobile money account at 2019-03-21 17:54:42. Message from sender: 1. Your new balance: 1301.34
SZL.Financial Transaction ID: 2*****3." toa="null" sc_toa="null" service_center="+268xxxxxxx" read="1" status="-1"
locked="0" date_sent="1553183682000" sub_id="-1" readable_date="21 Mar 2019 17:54:58" contact_name="(Unknown)" />
3 Services
4 <sms protocol="0" address="M-Money" date="1553188615580" type="1" subject="null" body="Your payment of 17.00 SZL to MTN
Bundles with token has been completed at 2019-03-21 19:16:21. Your new balance: 1284.34 SZL. Fee was 0.00 SZL. Message:
-. Financial Transaction ID: 2*****1." toa="null" sc_toa="null" service_center="+268xxxxxxx" read="1" status="-1"
locked="0" date_sent="1553188581000" sub_id="-1" readable_date="21 Mar 2019 19:16:55" contact_name="(Unknown)" />
5 Cash-out/Cash-in
6 <sms protocol="0" address="M-Money" date="1553412682929" type="1" subject="null" body="You xxxxxx xxxxxxxxx
(268xxxxxxx) have via agent: *****, withdrawn 380.00 SZL from your mobile money account: ***** at 2019-03-24
09:31:04 and you can now collect your money in cash. Your new balance: 0,34 SZL. Fee paid: 12.00 SZL. Message from
agent: -. Financial Transaction ID: 2*****7." toa="null" sc_toa="null" service_center="+268xxxxxxx" read="1" status="
-1" locked="0" date_sent="1553412664000" sub_id="-1" readable_date="24 Mar 2019 09:31:22" contact_name="(Unknown)" />
7 Gifted services
8 <sms protocol="0" address="M-Money" date="1553530939115" type="1" subject="null" body="Y'ello. You have received Airtime
worth SZL 10.00 SZL from 268xxxxxxx. Transaction ID 2*****9. If you are not registered on MTN Mobile Money, please
carry a copy of your ID to the nearest MTN agent." toa="null" sc_toa="null" service_center="+268xxxxxxx" read="1" status="
-1" locked="0" date_sent="155353092000" sub_id="-1" readable_date="25 Mar 2019 18:22:19" contact_name="(Unknown)" />
9 Promotional
10 <sms protocol="0" address="MobileMoney" date="1553849187339" type="1" subject="null" body="Y'ello customer, start
sending money via MTN Mobile Money and qualify for Likhandlela Insurance. Simply send YES to 2030 for free to qualify. Ts
&Cs apply. " toa="null" sc_toa="null" service_center="+26876011033" read="1" status="-1" locked="0" date_sent="
1553849182000" sub_id="-1" readable_date="29 Mar 2019 10:46:27" contact_name="(Unknown)" />
11 Bills
12 <sms protocol="0" address="M-Money" date="1554362960093" type="1" subject="null" body="Your payment of 98.00 SZL to SEC
with token 2*****91931*****8 has been completed at 2019-04-04 09:29:01. Your new balance: 0.87 SZL. Fee was 1.47
SZL. Message: -. Financial Transaction ID: 2*****1." toa="null" sc_toa="null" service_center="+268xxxxxxx" read="1"
status="-1" locked="0" date_sent="1554362941000" sub_id="-1" readable_date="4 Apr 2019 09:29:20" contact_name="(Unknown)"
/>
13 Loan acquisition.repayment
14 <sms protocol="0" address="MTN MoMo" date="1574695651981" type="1" subject="null" body="Your payment of 108.00 SZL to
MOMO LOANS with token has been completed at 2019-11-25 17:26:00. Your new balance: 42.33 SZL. Fee was 0.00 SZL.
Message: -. Financial Transaction ID: 3*****7." toa="null" sc_toa="null" service_center="+268xxxxxxx" read="1" status="
-1" locked="0" date_sent="157469556000" sub_id="-1" readable_date="25 Nov 2019 17:27:31" contact_name="(Unknown)" />
15 Notifications
16 <sms protocol="0" address="MoMoLoans" date="1574880346851" type="1" subject="null" body="Y'ello! Your MoMo Quick Loan
application has been approved at a limit of SZL100.00. Terms and Conditions apply." toa="null" sc_toa="null"
service_center="+268xxxxxxx" read="1" status="-1" locked="0" date_sent="1574880337000" sub_id="-1" readable_date="27
Nov 2019 20:15:16" contact_name="(Unknown)" />
Line 16, Column 257 Tab Size: 4

```

Figure 4.1.2: Examples of SMSs in xml format from the 8 different classes

In all, a total of 1010 SMSs were labelled for this project. The excel file was exported to a tsv* file for easier manipulation.

4.1.3 Data Augmentation

Data augmentation is a technique used to compensate for the scarcity of data which might lead to a dataset that may not help achieve a robust model. Besides increasing the size of the dataset, augmentation also increases variety in the data which helps the model to better generalize. It easily applied in computer vision by resizing and rescaling images. In text, one way to achieve augmentation is to either change the length of the sentence or replace words with their relevant synonyms and the latter that was applied in this project.

The table below shows the keywords that were augmented with one of their accompanying synonyms.

Keywords	Synonyms (for augmentation)
payment	discharge, remittance
completed	finalized, concluded
received	gained, accrued
transferred	sent
cashout	pay-off
transaction	settlement
new	latest
withdrawn	removed
balance	remainder

Table 4.1.3: A table of keywords and their augment word equivalents

The augmentation was done in levels and at each level the model was trained and tested. For instance, the first level augmented only one word and the model was trained and tested. At the second level, a second word was augmented, and the model was trained and tested as well. This means that for this project the levels were 9 since we have 9 keywords. Invariably, the original dataset of 1010 SMSs doubled in size after each augmentation phase. After the final level of augmentation, the dataset had 517,120 (that is, 1010×2^9) SMSs of which 20% were Mobile Money SMS.

4.1.4 Data Occlusion

The last stage of the data manipulation was occlusion. Occlusion in machine learning is covering a feature that might make the data trivial as the model is being trained. Triviality in the data works against achieving a general model which means that the model might fail if it comes across new data. All instances of MTN MoMo, MoMoLoans, M-Money, MobileMoney were occluded in one dataset. In the end there were two datasets, one occluded and the other not occluded and training and test were performed on both.

4.2 Classification Model

A plain neural network with ReLU (rectified linear unit) was used for the classification of the SMSs. There were two models of varying depths and parameters which were trained on the different datasets as explained above. The model that performs better was picked and used for the system.

```
Model(  
  (l1): Linear(in_features=1000, out_features=100, bias=True)  
  (l2): Linear(in_features=1000, out_features=100, bias=True)  
  (l3): Linear(in_features=100, out_features=2, bias=True)  
)
```

Figure 4.2A: Shows model 1 with 3 layers

```

Model(
  (11): Linear(in_features=1000, out_features=1000, bias=True)
  (12): Linear(in_features=1000, out_features=1000, bias=True)
  (13): Linear(in_features=1000, out_features=100, bias=True)
  (14): Linear(in_features=100, out_features=100, bias=True)
  (15): Linear(in_features=100, out_features=2, bias=True)
)

```

Figure 4.2B: Shows model 2 with 5 layers

Both models were built and tested in Google Colab using Pytorch and Python's machine learning libraries and packages like NumPy, pandas, torch, Tokenizer, matplotlib, etc.

4.3 Text Analysis

4.3.1 Formatting and Cleaning

The original format of the SMSs which is XML was retained. Special characters like single and double quotations marks were removed from each and every SMS to prevent them from obstructing access to key information and analysis. This done through a simple python script which replaced all special characters with blanks. The script was run on Visual Studio Code.

4.3.2 Analysis

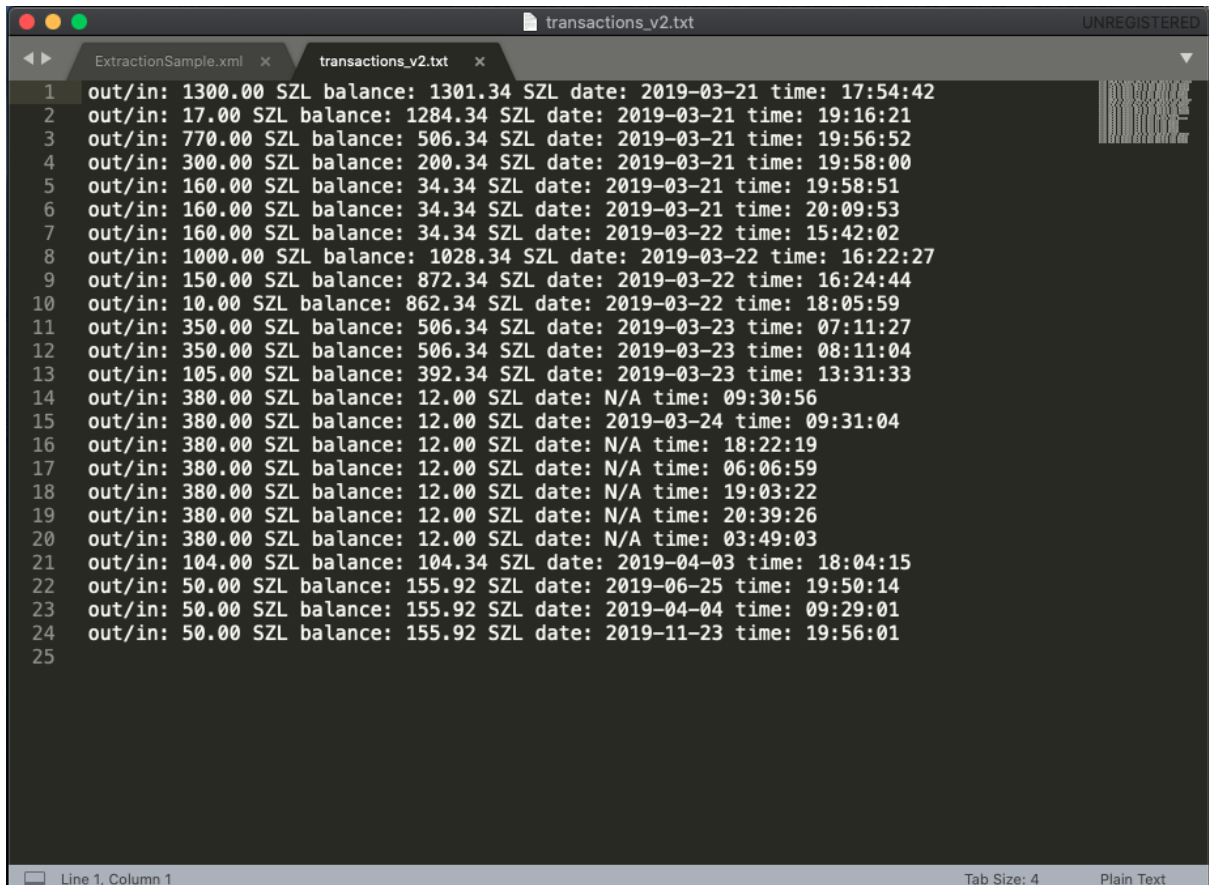
The analysis of the text was done in four steps:

Step 1: Identifying keywords that might point to key information like type of transaction, balance, etc.

Step 2: Extract the key information in some unique relation with the keywords. Once the keywords were identified, regular expressions (regex) patterns were created for all the information that needed to be extracted from each SMS. In all, there were three regex patterns for monetary values, date, and time. To implement the regex in python script, the RE package was imported and employed accordingly. The identified information was then

stored in list with each item in the list representing a single transaction with its key information.

Step 3: Formatting and writing the key information into a text file. The information was formatted in such way that it followed the conversion of financial statements.



```
1 out/in: 1300.00 SZL balance: 1301.34 SZL date: 2019-03-21 time: 17:54:42
2 out/in: 17.00 SZL balance: 1284.34 SZL date: 2019-03-21 time: 19:16:21
3 out/in: 770.00 SZL balance: 506.34 SZL date: 2019-03-21 time: 19:56:52
4 out/in: 300.00 SZL balance: 200.34 SZL date: 2019-03-21 time: 19:58:00
5 out/in: 160.00 SZL balance: 34.34 SZL date: 2019-03-21 time: 19:58:51
6 out/in: 160.00 SZL balance: 34.34 SZL date: 2019-03-21 time: 20:09:53
7 out/in: 160.00 SZL balance: 34.34 SZL date: 2019-03-22 time: 15:42:02
8 out/in: 1000.00 SZL balance: 1028.34 SZL date: 2019-03-22 time: 16:22:27
9 out/in: 150.00 SZL balance: 872.34 SZL date: 2019-03-22 time: 16:24:44
10 out/in: 10.00 SZL balance: 862.34 SZL date: 2019-03-22 time: 18:05:59
11 out/in: 350.00 SZL balance: 506.34 SZL date: 2019-03-23 time: 07:11:27
12 out/in: 350.00 SZL balance: 506.34 SZL date: 2019-03-23 time: 08:11:04
13 out/in: 105.00 SZL balance: 392.34 SZL date: 2019-03-23 time: 13:31:33
14 out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 09:30:56
15 out/in: 380.00 SZL balance: 12.00 SZL date: 2019-03-24 time: 09:31:04
16 out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 18:22:19
17 out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 06:06:59
18 out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 19:03:22
19 out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 20:39:26
20 out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 03:49:03
21 out/in: 104.00 SZL balance: 104.34 SZL date: 2019-04-03 time: 18:04:15
22 out/in: 50.00 SZL balance: 155.92 SZL date: 2019-06-25 time: 19:50:14
23 out/in: 50.00 SZL balance: 155.92 SZL date: 2019-04-04 time: 09:29:01
24 out/in: 50.00 SZL balance: 155.92 SZL date: 2019-11-23 time: 19:56:01
25
```

Figure 4.3.2A: Showing the text file with the key financial information that was extracted from SMSs

Step 4: Converting the text file into a portable document format, PDF. For this purpose, the FPDF package in python was used. The package has methods for reading information from a text file and formatting options before writing the information on the output PDF.

Username and ID + Cell Number

Financial Statement [dates]

out/in: 1300.00 SZL balance: 1301.34 SZL date: 2019-03-21 time: 17:54:42

out/in: 17.00 SZL balance: 1284.34 SZL date: 2019-03-21 time: 19:16:21

out/in: 770.00 SZL balance: 506.34 SZL date: 2019-03-21 time: 19:56:52

out/in: 300.00 SZL balance: 200.34 SZL date: 2019-03-21 time: 19:58:00

out/in: 160.00 SZL balance: 34.34 SZL date: 2019-03-21 time: 19:58:51

out/in: 160.00 SZL balance: 34.34 SZL date: 2019-03-21 time: 20:09:53

out/in: 160.00 SZL balance: 34.34 SZL date: 2019-03-22 time: 15:42:02

out/in: 1000.00 SZL balance: 1028.34 SZL date: 2019-03-22 time: 16:22:27

out/in: 150.00 SZL balance: 872.34 SZL date: 2019-03-22 time: 16:24:44

out/in: 10.00 SZL balance: 862.34 SZL date: 2019-03-22 time: 18:05:59

out/in: 350.00 SZL balance: 506.34 SZL date: 2019-03-23 time: 07:11:27

out/in: 350.00 SZL balance: 506.34 SZL date: 2019-03-23 time: 08:11:04

out/in: 105.00 SZL balance: 392.34 SZL date: 2019-03-23 time: 13:31:33

out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 09:30:56

out/in: 380.00 SZL balance: 12.00 SZL date: 2019-03-24 time: 09:31:04

out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 18:22:19

out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 06:06:59

out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 19:03:22

out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 20:39:26

out/in: 380.00 SZL balance: 12.00 SZL date: N/A time: 03:49:03

out/in: 104.00 SZL balance: 104.34 SZL date: 2019-04-03 time: 18:04:15

out/in: 50.00 SZL balance: 155.92 SZL date: 2019-06-25 time: 19:50:14

out/in: 50.00 SZL balance: 155.92 SZL date: 2019-04-04 time: 09:29:01

out/in: 50.00 SZL balance: 155.92 SZL date: 2019-11-23 time: 19:56:01

Figure 4.3.2B: Sample document showing a pdf document which was converted from a text file

Step 5: Embedding security features to the PDF. The first feature was a watermark which was used for the purpose of this applied project to present the various possibilities of securing a document. For this, the FPDF package was used. It took the watermarking image and overlaid it on the PDF document. The second security feature was a QR code which was generated using the user information: name, ID number and number. To generate the QR

code, the qrcode package in python was used and the FPDF package was used to overlay the code on the PDF document.

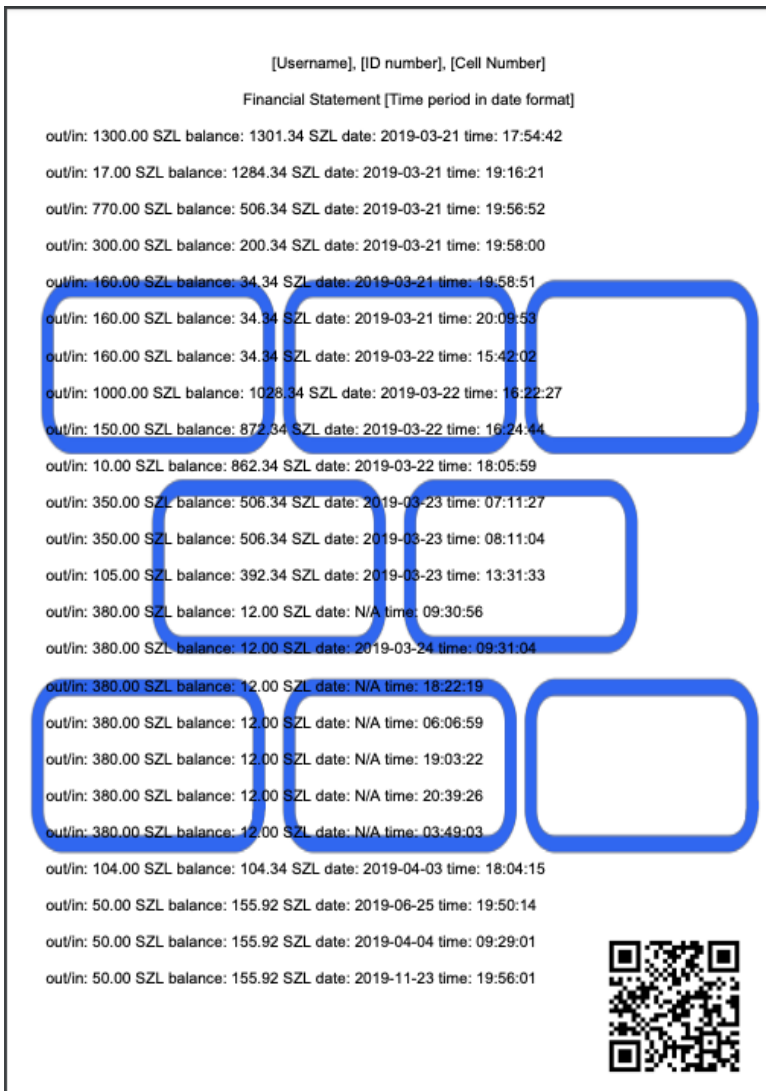


Figure 4.3.2C: Showing the pdf document with security features: watermark and QR code.

Chapter 5 - Testing and Results

5.1 Text Classification

Primary tests were done on the two models using bootstrapped training sets of 1010 SMSs which were augmented up to 6 levels. The fully augmented data was too large and took too long to train and was resource intensive because it uses the internet since Google Colab is a cloud service.

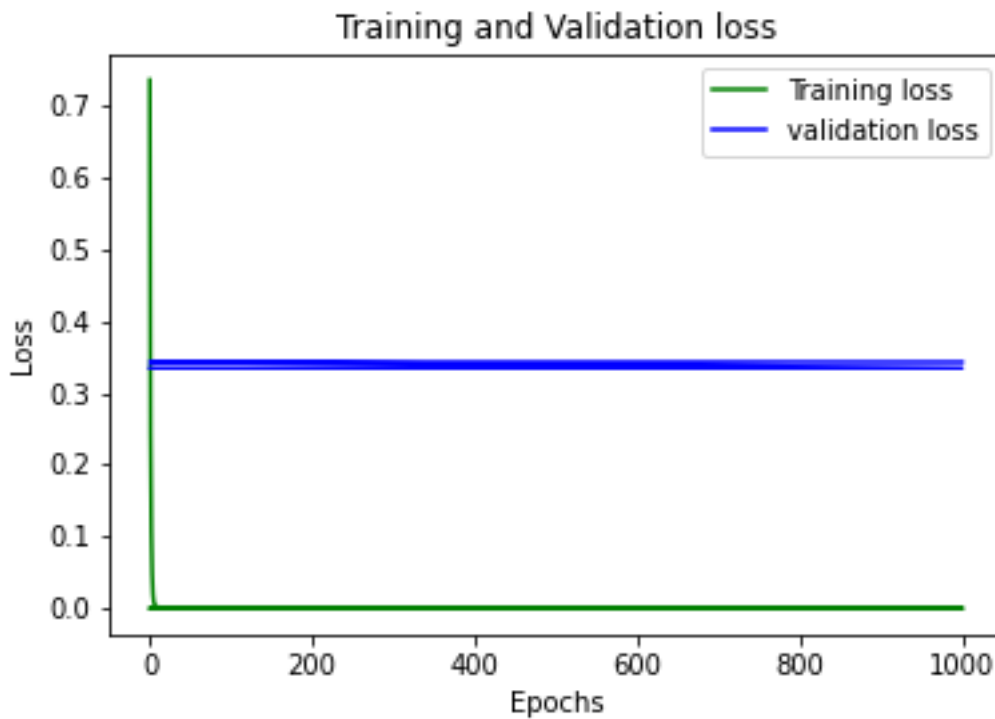


Figure 5.1A: Shows the training and validation loss for Model 1 on the original dataset of 1010 SMSs

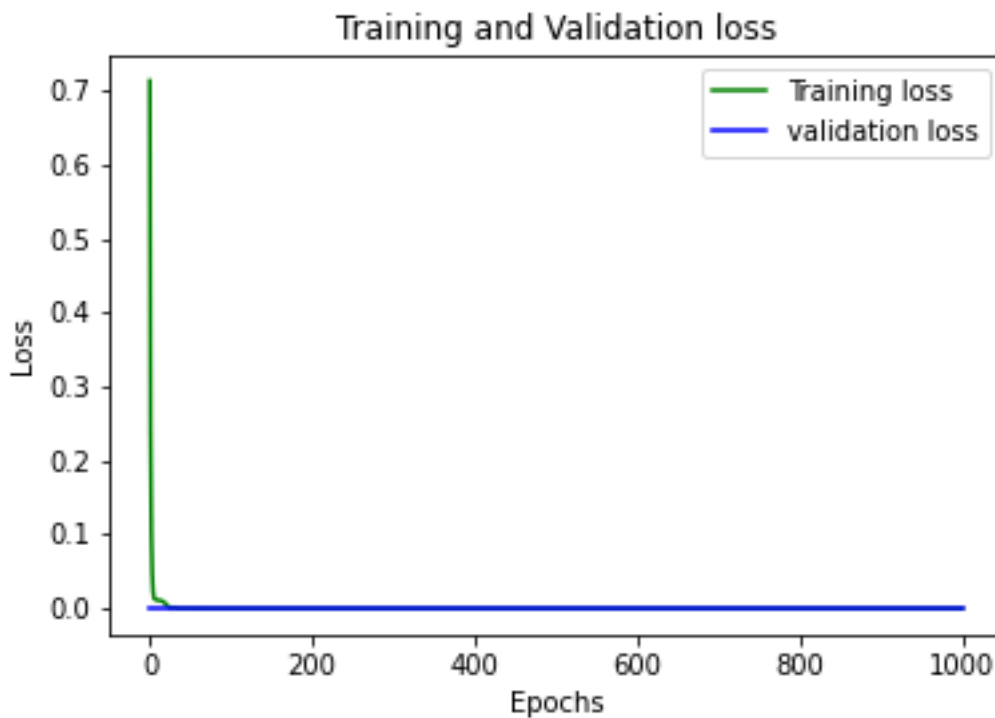


Figure 5.1B: Shows the training and validation loss for Model 1 on the augmented dataset of 32,320 SMSs

The figures above, Figure 5.1A and 5.1B, show the difference that augmented data can make in the performance of model. Augmented data produces a more stable and accurate model as shown by the loss of 0 in the model in Figure 5.1B when compared to the original data where the validation loss is above 0.3 yet the training loss is 0. Another thing these two graphs show is the necessity for early stopping in the training process since the loss of 0 on 1000 epochs is achieved before 100 epochs.

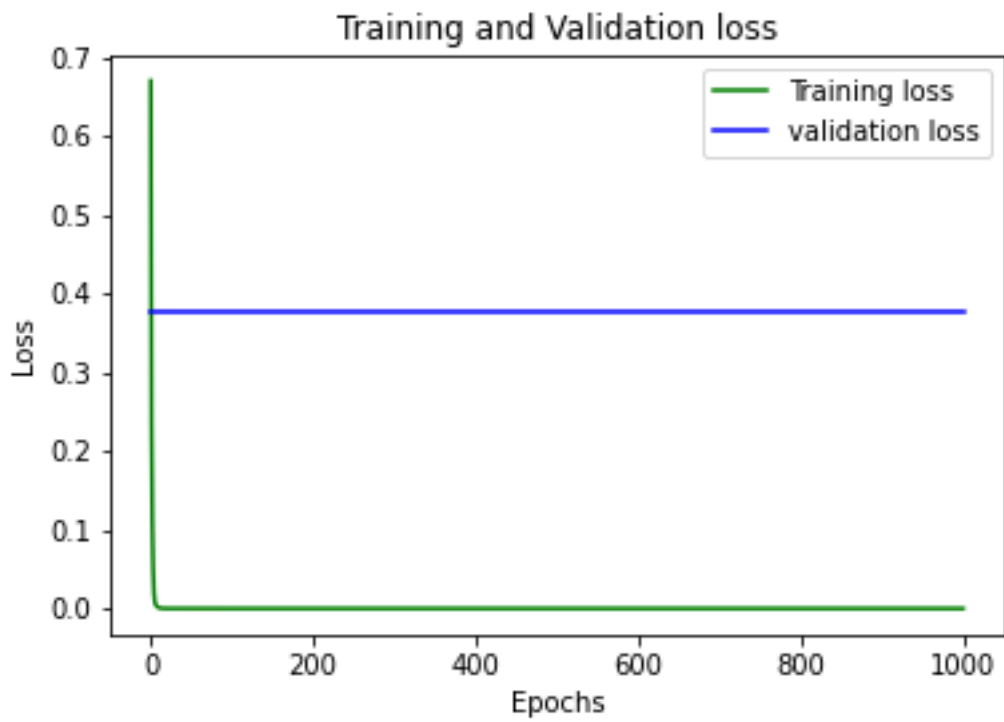


Figure 1.1C: Shows the training and validation loss for Model 2 on the original dataset of 1010 SMSs

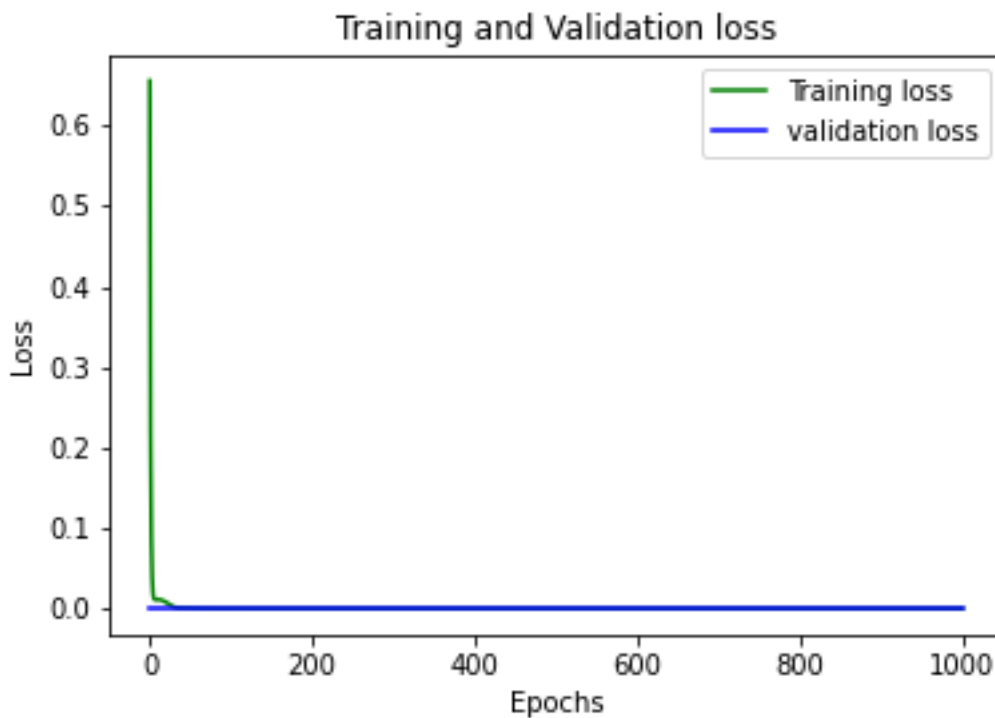


Figure 5.1D: Shows the training and validation loss for Model 2 on the augmented dataset of 32,320 SMSs

The same phenomenon which was present in the training of Model 1 is also present in Model 2 as shown in figure 5 and 6 above. The only difference is the slightly higher validation loss on the original dataset, again, putting emphasis on the need for augmented or a diverse dataset.

Interestingly, the depth of Model 2 does not better the performance of Model 1. This suggests that there are no complex features on the data which might be explained by the nature of the data since its semi-structured. On this regard it might be good to go with Model 1 since it will be small and more portable compared to Model 2.

5.2 Text Analysis

Testing on this part would have required presenting the financial statements produced by the system to banks and getting their feedback on the information presented and the security

features in the document. Low level testing included testing the system when the number of pages for each financial statement increased to more than one page and it was handled well. Another test included adding junk SMSs to the classified mobile money SMSs. Even though some of these junk SMSs have some key information that is also in the Mobile Money SMSs, the lack of completeness of this information prevents junk information from being inserted in the list of transactions. What remains to be tested and was not tested here due to the lack of such data, is how the system would handle SMSs which contain financial information like those from the bank. Such information if added together with the Mobile Money transactions might distort the financial statement.

Chapter 6: Conclusions and Future Work

The possibilities that come with analyzing SMSs for enhancing financial inclusion are endless especially when machine learning is exploited in classification stages of this process. For instance, in this applied project the key financial information that was stored in a list could be used as a secondary data source for predicting things like loan repayments for users. The critical thing which this applied project sought to contribute is the approach to such innovation problems. The approach as outlined earlier should ask nothing more from the user than they already have, should adopt a fluid digital footprint for user convenience, and the services offered by integrated platforms should be dynamic and take into account the activities of the user in other platforms. The system which was presented in this applied project was very basic and the idea was to help indicate what's possible even for populations whose digital footprint is limited. Further work might include expanding the scope of the SMSs beyond mobile money SMSs but that would probably also shift the focus from the unbanked populations whose digital footprint is limited. Second, this work was limited by the lack of bridge technology between machine learning and feature phones, which are prevalent with the unbanked population. Future work on this front might include the exploring ways by which machine learning can benefit the users of feature phones especially in respect to financial inclusion. Another possibility is using the extracted information to extend the information by inspecting how each transaction relates to the next especially in terms of dates, and amounts received and transferred. After that, maybe there might each users' financial statement might be graphed and predicted accordingly. By exploring the limited digital footprint that unfortunate groups like the unbanked population have, financial inclusion can be enhanced in many interesting ways.

References

- [1] Akomea-Frimpong, I., Andoh, C., Okudzeto, A., & Dwomoh-Okudzeto, Y. (2019). Control of fraud on mobile money services in Ghana: An exploratory study. *Journal of Money Laundering Control*, 22, 20–39. <https://doi.org/10.1108/JMLC-03-2018-0023>
- [2] Centellegher, S., Miritello, G., Villatoro, D., Parameshwar, D., Lepri, B., & Oliver, N. (2018). Mobile Money: Understanding and Predicting its Adoption and Use in a Developing Economy. *ArXiv:1812.03289 [Physics]*. Retrieved from <http://arxiv.org/abs/1812.03289>
- [3] Adedoyin, A., Kapetanakis, S., Samakovitis, G., & Petridis, M. (2017). Predicting Fraud in Mobile Money Transfer Using Case-Based Reasoning. In M. Bramer & M. Petridis (Eds.), *Artificial Intelligence XXXIV* (pp. 325–337). Springer International Publishing.
- [4] Linda Du. (2019, January 30). Can Mobile Money Boost Financial Inclusion in Southern Africa?. Retrieved October 14, 2019, from Yale Insights website: <https://insights.som.yale.edu/insights/can-mobile-money-boost-financial-inclusion-in-southern-africa>
- [5] Cook, T., & McKay, C. (2015, April). *How M-Shwari Works: The Story So Far*. 24.
- [6] Blechman, J. G. (2016). Mobile credit in Kenya and Tanzania: Emerging regulatory challenges in consumer protection, credit reporting and use of customer transactional data. *The African Journal of Information and Communication (AJIC)*, 17, 61-88.
- [7] UFA2020 Overview: Universal Financial Access by 2020. (n.d.). Retrieved October 13, 2019, from <https://www.worldbank.org/en/topic/financialinclusion/brief/achieving-universal-financial-access-by-2020>
- [8] National Financial Inclusion Strategy for Swaziland 2017-2022. (n.d.). Retrieved October 14, 2019, from Alliance for Financial Inclusion | Bringing smart policies to life website:

<https://www.afi-global.org/publications/2680/National-Financial-Inclusion-Strategy-for-Swaziland-2017-2022>

- [9] “Times Of Swaziland.” <http://www.times.co.sz/business/122003-mtn-sbs-launch-escrow-account.html> (accessed May 08, 2020).
- [10] T. Chen and M.-Y. Kan, “Creating a Live, Public Short Message Service Corpus: The NUS SMS Corpus,” *Lang Resources & Evaluation*, Aug. 2012.
- [11] “GSMA | The Mobile Economy Sub-Saharan Africa 2019 - GSMA Sub Saharan Africa.” <https://www.gsma.com/subsaharanafrika/resources/the-mobile-economy-sub-saharan-africa-2019> (accessed May 08, 2020).
- [12] Alexander, Alex J., Lin Shi, and Bensam Solomon. “How Fintech is Reaching the Poor in Africa and Asia: A Start-Up Perspective”, EM Compass Note 34, IFC; Saal, Matthew, Susan Starnes, and Thomas Rehermann. “Digital Financial Services: Challenges and Opportunities for Emerging Market Banks”, EM Compass Note 42, IFC.