CATHOLIC UNIVERSITY OF BUKAVU

B.P.285 BUKAVU

FACULTY OF SCIENCES

Department of Computer Science

Development of a messaging application for communication and detection of spam on a mobile operator, case study of Airtel, Vodacom and Orange.

Presented by : **MURHULA BYABUSHI Christian**

Dissertation presented and defended in order to obtain the degree of Bachelor in Computer Science.

Option: Network and Telecommunications
Degree: Final year of Bachelor

Supervised by: ***Hw.*** **MUGISHO MUSHEGERHA Youen**
Directed by : ***PhD.*** **Elie ZIHINDULA**

Academic year: 2022-2023

Contents

List of Figures

List of Tables

Introduction

0.1 Context and generalities

With the increasing of use of mobile devices in mobile telecommunication, the number of text messages sent every day has grown exponentially. According to *Statista*, a company that provides market and consumer data on a wide range of topics, including digital media and technology; the number of mobile messages sent worldwide in 2020 reached 3.5 trillion [31]. In the same case, with the raise of web pages and social media messaging applications like Whatshap, Teelgra, Snapchat Facebook, Instagram and many others, phone users can now send messages that are only based on text as in former time but also on video, audios which are more chestful comparatively [14].

For sure, for interacting with his partner more professionally, email message is the most mean used, but not in all cases since in some countries a given SIM Card of a telecommunication provider is used as a bank more than being an communication mean, so the importance to secure the communication lines of a mobile users in these countries. Along with the functions and interests that the mobiles messages encompasses in terms of conversing, money sending and receiving, there has been an increase in the number of spam messages that aim to deceive people into providing personal information, sending unwillingly money, menacing to death or taking other actions that benefit the scammer.

To address this problem, the development of a messaging application with advanced spam detection capabilities is crucial to set, filter messages and prevent the users. This dissertation focuses specifically on the development of such application for phone users and in general for mobile networks telecommunications technologies.

0.2 Problematic

In telecommunication domain, we use mobile devices or phones for sharing *SMS*, Email, chats by using some specific apps. Among all we use specifically *SMS* for personal and professional information sharing [24]. The SMS stands for Short Message Service, which is a text messaging service for mobile phones and other mobile devices. It allows users to send and receive short messages of up to 160 characters [25]. It is also possible to send or receive automatic SMS which are not sent by human, whereas by using web interface or API [21].

Time to time, more persons are receiving messages such as : "You won x amount of money send another amount to withdraw it", "Join me at x area to take your money but pay me the transport ...", "I'm *Sirene* Madam I have money for you", "I have a job for you" and many others fake messages. More of them are reported in this spams examples

link.

Furthermore, scammers go the point where they can introduce vulnerabilities in messages which exploit a weakness in the SMS messaging system to remotely install spyware on mobile devices commonly called *Simjacker* [8]. However, in marketing almost similar messages are used to sensitize people to by products and services which confuse users wether it is or not a spam message by leading users to ignore important messages or being more hesitant to engage with mobile marketing campaigns [7],[27].

Considering all above issues caused by spam, what are key technical challenges that could be addressed in messaging system which can effectively facilitate the communication as well as detecting and filtering spam messages ?

0.3 Hypotheses

According to the Oxford dictionary, hypotheses stands for a statement of the expected relationship between things being studied, which is intended to explain certain facts or observations [32]. An idea to be tested. Hence, using the content-based filtering techniques, which involves analyzing the content of messages and determining wether it is a spam or not would be considered as solution.

Firstly this would be done by utilizing the Machine learning algorithms which are : **Naive Bayes, Logistic Regression, and Supper vector Machines**. All these would be combined by the ensemble methods for making a more predictive model [33] including the preparation and classification techniques which avoid biased model [22].

Secondly, during the production steps, we should integrate the model inside of the system able to add technically in a blacklist or whitelist suspect users based on the specific probability of being a potential attack.

Overall, we will jump from Machine Learning as model (*MLaaM*) which is the output of writing ML algorithms run on data and represent what was learned by the algorithm on training data; to ML Model Software Deployment which encompasses all the activities that make a software system worthy to be used [17].

0.4 Delimitation and objectives

0.4.1 Delimitation

The present work aims to develop a messaging application for communication and detection of spam on a mobile network.

Geographically it focuses on all provinces of Democratic Republic Of Congo(DRC) where mobile phones are used and require techniques for implementation.

Besides, it does not function effectively across all languages unless the solution model has been specifically trained on those languages. As a result, it is challenging to claim its effectiveness in languages such as Swahili, Lingala, French, or even English. Moreover, achieving optimal performance often necessitates the involvement of a large population.

Indeed, the current work in terms of planning and execution has spanned a duration of nine months : From January to November 2023.

0.4.2 Objectives

This system present 2 types of objectives such as: functional and non functional.

As functional objectives, this system consist of developing a messaging application that can facilitate efficient communication between users; implementing the machine learning models which has the capacity of classifying the messages and preventing spam messages under a certain probability; designing and integrating a user-friendly interface for messaging application;

For non-functional objectives, it allows the user whose messaging application complies with relevant privacy laws and regulation to protect data information, reducing the attacks and frauds; Increasing the trustfulness of users and mobile services compagnie provider, optimizing the power resources of the user against threats posed by scammers.

0.5 Interest

Personally, this paper has allowed the author to gain knowledge and more experience in the field of mobile networks and messaging applications.

Socially, the developed system contributes to facilitating communication and reducing the impact of spam messages, which can be annoying and stressful for citizens.

Economically, this system of detection helps to save business server time and resources by filtering out spam messages allowing for more targeted marketing efforts.

Scientifically, the research achieved contributes to the advancement of the science in the domain of Mobile Networks, SMS messaging, and Machine Learning data processing and classification.

0.6 Research Methodology

Throughout this paper, the research methodology will be used to guide the study towards achieving its objectives. The research will adopt a descriptive research design to describe the development of a messaging application for communication and detection of spam on a mobile network. The study will focus on both qualitative and quantitative research methods [10]. The qualitative method will involve a literature review, interviews and analysis of collected messages. while quantitative method will focus on the development and testing of the messaging application.

The research will be conducted in two phases. The first phase will involve data collection through a survey questionnaire that can be completed on website, or can be directly provided to the web interface (API) by the mobile phone users for collecting their experience with messages and especially spams. Thus, the data collected will be analyzed using descriptive statistics [4] to identify the common types of spam messages and the frequency of occurrence, languages inside, and other attributes.

The second phase will involve the development of the messaging application using the data collected from the survey and the analysis of existing messaging applications. The development of the application will be guided by the principles of agile software development by using Python (Django framework) for *Back-end* and HTML,CSS and JavaScript for *front-end*. Then,the application will integrate the use of machine learning models, including Naive Bayes, Logistic regression, and Support Vector Machine.

The evaluation of the messaging application will be conducted using both quantitative and qualitative methods. The qualitative evaluation will involve the measurement of the

application's accuracy and efficiency in detecting and filtering spam messages, on the other hand the qualitative evaluation will involve a user study to determine the usability and user experience of the application.

0.7 Work Plan (or Work Subdivision)

The work plan of this dissertation is divided into four parts. The first is the introduction, which provides a background information on the research problem. The second part consist of a situation analysis and assessment while the third part focuses on literature review and explanations on the methodology. Then the fourth part presents the practical result of this work. Finally the conclusion part summarizes the key findings and contributions of the study and presents limitations and provide recommendations for future research.

Chapter 1

Situation analysis and assessment on mobile phones

1.1 Introduction

In this chapter, we will focus on various aspects that enhance the comprehensiveness and practicality of this dissertation. It includes explanations of mobile messaging architecture, machine learning models, and spam messages in mobile world. Additionally, it provides an analysis of the architecture used by network operators, highlighting both positive and negative aspects of their approach to message handling.

1.2 Presentation of the working framework and definition of key concepts

1.2.1 Definition of key concepts

a) SMS(Short Message Service) :
The Short Message Service is a basic service allowing the exchange of short text messages between subscribers [26]. For supporting virtually all mobile devices, SMS is considered as a universal means of communication that enables users to communicate and function even though all users are not active simultaneously (asynchronous communication).

b) Enhanced Messaging Service (EMS) :
EMS has been created to allow the transmission of richer and more advanced messages. Unlike traditional SMS, EMS accepts not only text messages but also audios, melodies, and animations [25].

c) MMS (Multimedia Messaging Service):
MSS has been developed to facilitate the transmission of rich multimedia content in mobile messaging. Unlike SMS and EMS, MMS enables users to send not only text messages but also various types of multimedia files such as images, videos, audio recordings and even slideshows [25].

d) Spam message:
A spam message is understood as an unsolicited or undesired messages received on mobile phones which constitutes veritable nuisance to the mobile subscribers [34]. Clearly, this message can be sent with the intention of gaining financial benefits, collecting personal or organizational information such as security numbers, credit card details, or login credentials, and soliciting money by making false promises of future benefits or rewards that do not materialize.

e) Networks operators: The networks operators refers to companies or organizations that provide and manage telecommunication networks. These operators own and operate the infrastructure, such as mobile networks, fixed-line networks, or internet service provider (ISP) networks, that enable the transmission of user's information to another user of the network [15]

f) Artificial Intelligence (AI) : AI refers to the field of computer science that focuses on creating intelligent machines or systems that can perform tasks that would typically require human intelligence. For being practical, it encompasses algorithms, models and technologies that enable computers and machines to simulate human like cognitive processes such as learning, reasoning, problem-solving, perception and language understanding.

g) ML (Machine Learning) : Machine learning is a subfield of Artificial Intelligence that focus on the development of algorithms and models that enable computers to learn from data and make decisions or predictions without being explicitly programmed[36]. Clearly, when the data is labeled during the training, we refer to it as supervised model. If contrast, when the data is unlabeled and the model must discover patterns and relationships itself, it is an unsupervised model. Additionally, whenever it performs both the labeling and discovering patterns tasks, it refers to a semi-supervised model. Furthermore, there is the last type called reinforcement. This one, is used to teach a computer or an AI agent how to make series of decisions in an environment. Just like, we learn to play the game better by playing it over and over.

h) NLP (Natural Language Process): NLP is a subfield of Machine Learning that studies the human language and combing techniques from statistics, linguistics, life-hoods for making sentiment analysis, text classification, machine translation, question answering and text generation in a way that it can be understood computationally [6].

1.2.2 Presentation of the working framework

In the eastern party of DRC (South Kivu- and North Kivu) the usage of mobile phones has become more common, transforming communication and connectivity in the region. The DRC itself is large country, covering over 2,345,000 square kilometers with the eastern provinces of North and South Kivu spanning approximately 59,483 and 65,070 square kilometers respectively [39]. According to recent statistics from *GlobalEdge* [1], an Amer-

[1]GlobalEdge : Created in 1994 by the International Business Center and the Eli Broad College of Business at Michigan State University (IBC), globalEDGE™ is a knowledge web-portal that connects international business professionals worldwide to a wealth of information, insights, and learning resources on global business activities

ican company, around 95 million people were living in the DRC in 2022 [13], of which approximately 46.9% had active mobile phones based on *GSM* [2] research.

In this context, it is observed that more people in cities use mobile phones compared to those in villages, primarily due to limited accessibility. A research study conducted by *Target Canibet* [3]in 2015 focused on mobile connections in DRC cities including Bukavu, Goma, Kinshasa, Lubumbashi, and Matadi, found that among 1,000 people surveyed in each city, 9 out of 10 individuals were subscribed to a network operator. However, it was noted that approximately half of them subscribed to two operators, while a quarter subscribed to four operators, and 18% used the services of a single operator.

Furthermore, the recent statistics made by *DataReportal* [4] in DRC shows that the mobiles users continues to increase exponentially merely because of new services provided by internet and Telecoms Operators, at the point that since 2021 to 2022, it is has been reported 3.6 million of new users between 2021 to 2022, a report that proves how much mobile phones is inevitable in this last decades.

1.2.3 Network coverage and infrastructure

In fact, two telecoms services exist in DRC such as : Fixed services (26%) and Mobile services (74%). The first one known as landline or wired services, involve the use of physical infrastructure; the second one which is popular is the mobiles services refer to telecommunications services provided through mobile networks. According to the Congolese Regulatory Agency (ARPTC), the DRC has four mobile operators - Vodacom RDC, Airtel Congo, Orange DRC and Africell DRC. Vodacom is the leader in the voice segment, with 35.2% of the market, followed by Orange (30%), Airtel (23.9%) and Africell (10.9%). In the mobile internet market, Vodacom has 37.44%, Airtel 31.25%, Orange 28.14% and Africell 3.17% [3].

Additionally, since the 190s, when the DRC witnessed the first installation of operator systems such as Celtel(now Airtel) and Vodacom, followed by Orange and Africell, the telecomunications sector has shown significant market growth, reaching 1 Billion in 2022$ and expanding at a rate of 21% per year according to *GlobalData* [5]. However, this growth necessitates the updating of the infrastructure, which includes various generations of technologies, namely the second generation, third, fourth, and fifth(under development).

In fact, the second generation have been deployed in various territories to enable more efficient **voice calls, data networks services, and introduce SMS for text messaging**. The infrastructure required for 2G networks includes the following equipments: 1) BTS(Base Transceiver Station) : Transmit and receives signals between mobile devices. 2) MSC(Mobile Switch Controller) : serves as the switching entity that connects calls between mobile devices. 3) BSC (Base Station Controller) : manages multiples BTSs

[2]GSMA (Global System Communications Association) : An industry Organization which represents the interests of mobile network operators worldwide created in 1982 to ease cooperation between countries deploying *GSM* (Global System fo Mobile) technology.

[3]Target Canibet: Reseach & Consulting Group working in DRC. `https://www.target-sarl.cd/fr/content/etude-sur-la-telephonie-mobile-en-rdc`

[4]DataReportal: A online Company designed to help people and organizations all over the world to find the data, insights, and trends they need to make better informed decisions produced by Simon Kemp, `https://datareportal.com/reports/digital-2023-global-overview-report`

[5]GlobalData: Expert Company of Analysis, innovatove Solutions

and controlling radio resources, managing handovers between cells and optimizing network performance, 4) AuC (Authentication Center) responsible for managing subscriber authentication and encryption keys to ensure communication between mobile devices and the network, 5) Home Location Register (HLR) the database that stores subscriber information such as phone numbers, authentication details, and service profiles, 6) Visitor Location Register (VLR) : The VlR is a temporary database that stores information about roaming subscribers within a specific area 7) MS (Mobile Station), including all the technologies used by the users's handset and has two parts : **Firstly, the mobile equipment which contains the radio equipment, the user interface, the processing capability and memory requirements for call signaling, encryption, SMS and the id of the mobile phone(equipment IMEI number). Secondly the Subscriber Identity module (SIM Card),** used in encryption of codes needed to identify the subscriber, storing subscriber's information, locate the user [11] as (+243 for each congolese number).
Indeed, all the 2G technologies covers a large distance varying between 1880MHz - 2700 MHz.

Besides, the third generation appears as revolution, **allowing multimedia messages, voice calls data, faster data speed**; however it requires a significant upgrade from the previous generation. Thus, the equipment involved in 3G technology includes : 1) BTS (Base station Transceiver) : Which plays the same role as for 2G; 2) Node B: Responsible for handling the radio interface and connecting mobile devices to the core network; 3) Radio Network Controller (RNC) : Controlling the Node B and managing the radio resources 4) Mobile Switching Center (MSC): The MSC is the central switching entity in the network that **connects calls between mobile devices**; 5) Serving GPRS Support Node (SGSN) : Responsible for managing packet-switched data services services and handling mobility for mobile internet access; 6) Gateway GPRS Support Node (GGSN): It serves as interface between the mobile network and external networks like internet; 7) The Home Location Register (HLR) and Authentication Center(AuC): plays the same role as in 2G; 8) Operations Support System (OSS) : It provides and functionalities for monitoring and managing the 3G network. Indeed, the 3G is appreciated for enabling higher- speed services and covers different frequency bands depending on countries, ranging between 850Mhz - 1700 Mhz [29].

Additionally, the fourth generation, commonly referred to as LTE(Long-Term Evolution) represents a significant advancement over previous generations in terms of infrastructure and services. This generation introduces higher data speeds, improved capacity, and better perfomance for mobile communication and data services. The upgrades in infrastructure include: 1) BTS and MSCs : These components remain unchanged from the previous generation 2) Evolved Packet Core (EPC) The EPC is a critical component of the 4G core network architecture which provides the packet-switched backbone that handles data traffic and ensures efficient data delivery between mobile devices and the internet or other networks; 3) Radio Access Network (RAN): The Ran is responsible for the radio interface between mobile devices and base stations; 4) LTE (Long-Term Evolution) : is the primary air interface enabling the high data speeds, low latency; 5) Back-haul Network : It connects base stations to the core network and internet infrastructure; 6) Spectrum Allocation: Hands over the mobile operator access to specific radio frequency bands; 7) Network Management System: These systems monitor and manage the 4G

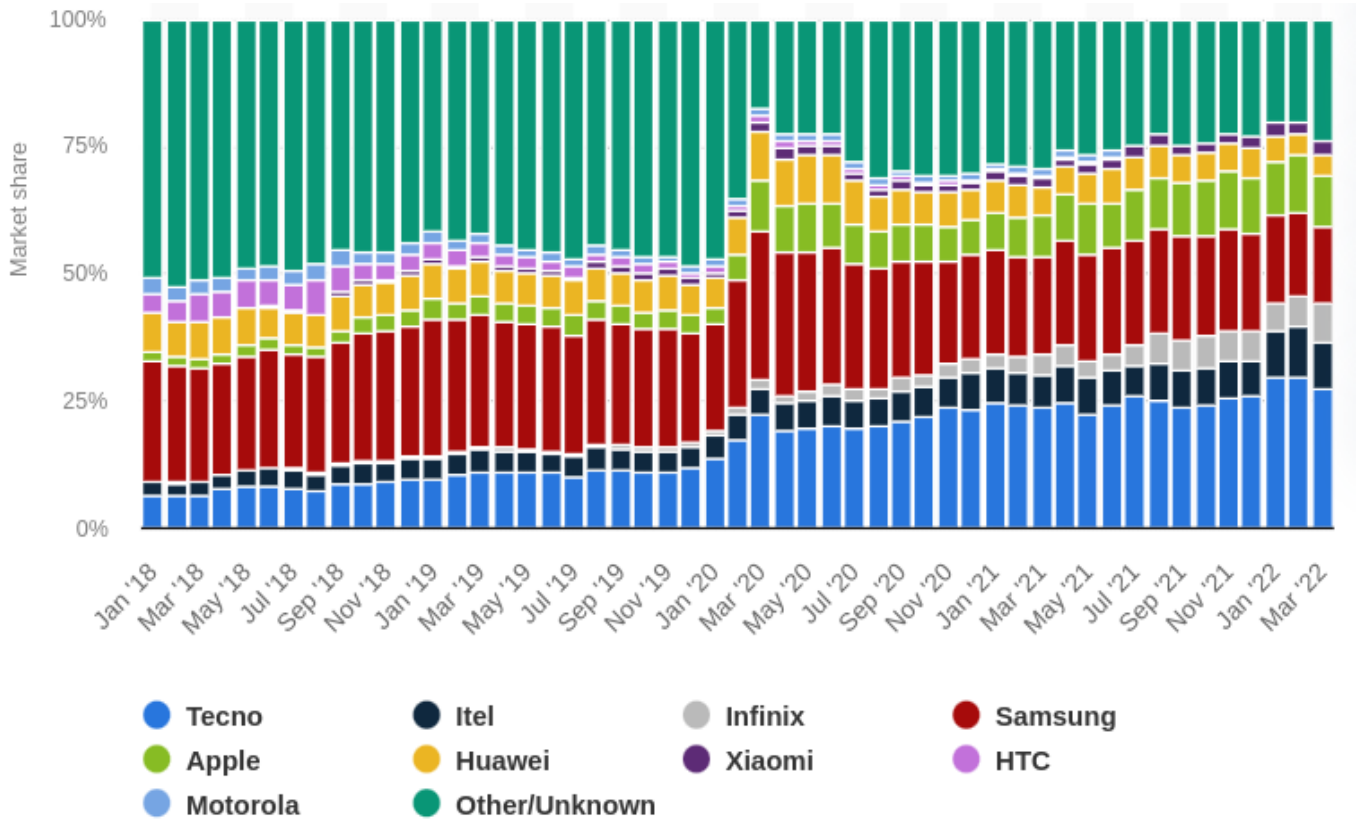

Figure 1.1: Market share of mobile device vendors in the Democratic Republic of the Congo from January 2018 to March 2022

network, ensuring its smooth operation, performance optimization, and troubleshooting. However, it's spanning or coverage of 4G networks on frequency bands allowed in each country based on their preferences, ranging from 700MHz to 2600 Mhz. The higher frequency bands generally offer faster data speeds but may have a shorter range, while the lower frequency bands can provide broader coverage but with slightly lower data speeds [29].

1.2.4 Mobile Phone Models

Since the mobiles phones are essential tools for communication, there is a wide range of popular mobile widely used by citizens of the DRC. The popular models come from various brands and offer a range of features to cater to different user preferences and needs. Some of the popular mobile phone models in DRC include *Techno, Itel, Infinix, Samsung, Apple, Huwaei, Itel, HTC, Motorola.* As it can be seen on the figure 1.1, according to the recent statistics conducted by *Statistica*, Samsung was the market leader in terms of share from January 2018 to November 2020, but in 2022, Tecno has emerged as the market leader.

Furthermore, all these models provided above sell the telephone following different types which include mobile phones, offering the features such as touchscreen displays, cameras, internet, connectivity, and access to mobile apps; features phones, which are basic mobile phones used for calling and texting; smartphones used by the majority (around 35% in DRC), providing access to mobile internet, mobile apps, multimedia messaging, and various productivity tools; Tablets, used for reading and for the same

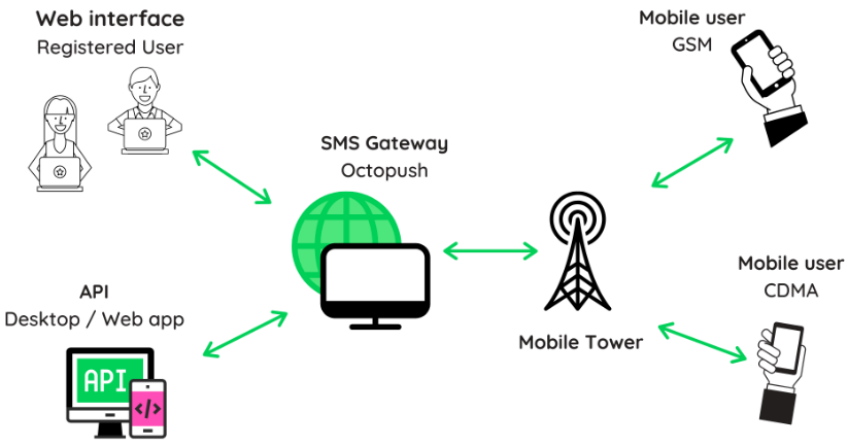

Figure 1.2: SMS Gateway Provider Architecture

functionalities as smartphones.

1.2.5 Mobile usage and prevalence

In fact, each phone has its own unique set of characteristics that define its capacity and performance compared to others. Some phones come with specific applications that can be used independently, even without being connected to an operator, such as a camera, calculator, games app, and many others. However, other phones may not have such features.

To access the services provided by the operator, the phone's sim-card must be functioning, recognized by the operator, and capable of sending and receiving communication signals. **Of course, all these services work only if the phone's battery has sufficient power**.

Moreover, the services that citizen's subscribers benefit from are as follows :

Firstly, the Internet Access: The Internet services are used to connect people from different nodes. In fact, In accordance with the *WorldBank* [6] been used by 23% of the DRC's population in 2020. Nonetheless, it requires payment which proportionally gives mobile data usually expressed in Megabytes.

Secondly, the Text Messaging (SMS): Even though the internet is the most used for texting, the SMS remains a widely used form of communication, especially in regions with **limited internet connectivity or among users who prefer simple text messaging or do not have the internet connection**. Furthermore, with the architecture of GSM(Global System for Mobile Communications) invented in the second generation, sending messages became possible. Nowadays, the web environment has developed application interfaces (API) that connect external systems to operators for sending messages [19]. One of the platforms that offer these services connects its SMS gateway to the GSM operator, as seen in the case of *Octopush* [7] architecture shown in Figure 1.2

Thirdly, the Mobile Banking and Payments : With this services subscriber can make **financial transactions**; paying bills, transferring money conveniently.

[6]WorldBank : International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database `https://data.worldbank.org/indicator`

[7]Octopush : SMS platform for businesses connected with their audience

Fourthly, the Mobile Entertainment: Mobile phones offer a range of entertainment options, from streaming videos and music to mobile gaming.

And finally, the others Mobiles apps: This party includes the health services, education, social medias applications.

1.3 Purposes of spammers in mobile messages

Most of the time, spammers prefer to promise recipients prizes and then ask for money to claim the offer. They also attack SMS gateways with DoS (Denial of Service) messages [1] whose goal is to overwhelm the system with unnecessary messages. Spammers send advertising and promotional messages based on company objectives, as well as SMS containing fake links or impersonating organizations to deceive recipients into taking certain actions or providing sensitive information [37]. Additionally, they may send SMS disguised as surveys to gather personal information for various fraudulent purposes.

1.4 Solutions

To address theses problems, it is necessary to involve various stakeholders, including network operators, app developers, regulatory bodies, and users. Firstly, it is recommended to implement mechanisms at the network level [18] to filter messages and block users involved in spamming. Secondly, users (subscribers) should be educated on how to analyze messages and report any one that is causing disturbances. Thirdly, regulatory measures should be enforced to establish stringent regulations and penalties for spammers and those engaged in fraudulent mobile activities. Fourthly, the development of apps that enable filtering, classification, and reporting in the subscriber side would be beneficial. Fifthly, a website can be set to collect messages whether spam or ham reported by users who have doubts about their legitimacy, and then Machine Learning models can be used for detection purposes. For this, supervised or unsupervised methods can be employed to classify and predicting whether a message is spam or ham.

1.5 Summary

Overall, with the growth of mobile technologies, subscribers benefit from diversified services including : text messaging, voice calls, mobile baking, entertainment apps, and many others. However, These advancements also bring new challenges, such as the development of spam messages that aim to disturb network subscribers with unwanted or threatening messages.

In DRC, especially in the eastern party, users face similar issues. This chapter emphasizes methods or techniques that can be used to address this problem and relatively reduce spam. One of the prominent techniques suggested is based on Artificial Intelligence, particularly Machine Learning algorithms.

Chapter 2

Review of the Literature and description of the approach

2.1 Introduction

This chapter delves into theory, methodologies, and machine learning techniques, including relevant algorithms and their deployment in the suggested solution. It also highlights the contributions of previous researchers in the field.

2.2 Revue of the Literature

Numerous researchers have extensively explored the subject of spam detection. Within this domain, some have directed their investigations towards the web environment, while others have delved into realm of mobile technologies.

Furthermore, these researchers have chosen to investigate the detection of spam across various communication channels, including emails and SMS, encompassing Multimedia SMS (MSS) as well. In the following sections, we comprehensively review the body of work that has been accomplished within this context as follows:

[23]. **Dr.V.M Veena K.Katankar** proposed a system that comprises an SMS gateway for transferring SMS messages after they have been stored and encrypted by the web server. This software operates through a web interface. Whenever a client sends a *POST* request, it is received by the web server, which is responsible for encryption or decryption if necessary. Subsequently, the gateway transfers the message as per its designated route. This solution proves to be particularly valuable in mobile banking and organizational marketing systems. Nevertheless, the author encourages other researchers to delve into channel services in communication and advanced encryption techniques.

[5] In their publication titled ***Short Message Service***, **Brown, Jeff and Shipman** members of IEEE, delve into several significant aspects. They start by exploring th growth of mobile phones and SMS services. They also examine the system architecture of SMS Centers and technologies used for message communication

Furthermore, they shine the spotlight on aggregators and services providers. These are the entities that enable users to send bulk messages, essentially sending messages with a large amount of text to a group of recipients. This includes the interesting capability of converting email messages into SMS.

Moreover, the article highlights that some of these aggregators may choose to collaborate with cellular networks. In this collaborative role, they act as intermediaries, connecting third-party entities that don't have direct relationships with cellular service providers. To achieve this, they employ a *SMPP (Simple Messaging Peer to Peer)* protocols.

[28] **Researchers A. Medani and A. Gani**, affiliated with the University of Malaysia, have published a comprehensive review focusing on security concerns and techniques related to mobile Short Message Service (SMS).

In their paper, they illuminate the process by wich a subscriber sends a message to another party while adhering to specific principles of the Over-The-Air (OTA) structure. This process involves transmitting the message from the sender to the base station and then forwarding it to the intended recipient trough the SMS Center (SMSC).

Crucially, they emphasize the importance of securing every SMS using *Public Key Infrastructure (PKI)*, ensuring end-to-end transmission security and safeguarding the message from unauthorized modifications. However, it's worth noting that the use of PKI can potentially impact mobile device performance due to the significant power requirements for the encryption process, and it may not guarantee integrity across all standards.

To address these security concerns within GSM systems, the researchers propose the implementation of *XML Key Management Specification* as a middleware solution. This middle ware system servers an intermediary, facilitating secure communication between mobile devices and enhancing overall system security for the benefit of clients.

[9] **Nikhil Kumar**, a reseacher affiliated with the University of New Delphi in India, published an article focusing on the topic of Email Spam Detection Using Machine Learning. In his study, he placed particular emphasis on comparing various machine algorithms, including ***Naive Bayes, Support Vector Classifier, AdaBoosting, K-Nearest Neighbour, and Bagging Classifier***. The objective was to predict wether an email was categorized as spam or legitimate (ham). To demonstrate his approach, he utilized an existing dataset available in the Kaggle workspace.

Through experimentation and parameter tuning, Nikhil found that Naive Bayes delivered promising results in terms of accuracy. However, he also pointed out a limitation associated with the Naive Bayes algorithm. This limitation is tied to its assumption of class-conditional independence, which implies that each feature is considered independent of the presence of the other features. In cases where this assumption does not hold, it can lead to misclassification of data points.

To address this limitation and enhance the performance of spam detection, the author recommended the use of **ensemble methods**. These methods involve the use of multiple classifiers for making class predictions, allowing for more robust and accurate results.

[30] In 2018, researchers Pavas Navaney, Gaurav Dubey, and Ajax Rana, who are affiliated with the University of Southern California and Amity presented a conference paper titled "SMS Spam Filtering using Supervised Machine Learning Algorithms".

Their study concentrated on a dataset comprising 5574 records, of which 4827 messages were categorized as "ham" (legimate messages), while 747 messages were classified as "spam" (unsollicited or unwanted messages).

The researchers applied three different machine learning methods to this dataset. Among these methods, it was observed that the *Support Vector Machine (SVM)* algo-

rithm achieved the highest accuracy compared to the Naive Bayes and Maximum Entropy Classifier algorithms.

[35] **Houshmand Shirani-Mehr**, a researcher in Machine Learning, published an article in 2013 titled : "SMS Spam Detection using Machine Learning Approach". His purpose was to address the spam filtering problem by utilizing ML algorithms. Therefore, He utilized a dataset from the *UCI Machine Learning Repository* repository [1] , which contained real SMS messages. In development, he employed the algorithms to tackle that problem such as : Naive Bayes with Laplace smoothing, Support Vector Machine, and Ensemble methods (*AdaBoosting and Random Forest*). As an improvement, the author added meaningful features such as the length of messages in terms of the number of characters and certain thresholds.

The results obtained after applying these methods to the dataset indicate that the SVM algorithm achieved the highest accuracy score.

[16] The authors of the article titled "SMS Spam Detection Using Machine Learning", namely **Gupta, Suparna Das and Saha, Soumyabrata and Das, Suman Kumar**. Their focus was on reviewing various techniques employed by other researchers in the realm of machine algorithms for SMS spam detection.

In their research, they adopted a similar approach by incorporating the *TF-IDF (Term Frenquency-Inverse Document Frequency)* method. This technique assesses the frequency of a word within a document and evaluates its importance in that document. *TF-IDF* is a well-known method for measuring word relevance in a collection of texts.

To assess the effectiveness of these techniques, the authors applied them to a spam dataset obtained from *Kaggle* [2] . After conducting their experiments and evaluations, the authors arrived at a noteworthy conclusion. They found that among all the ML algorithms they employed, the Naive Bayes algorithm consistently achieved the highest level of accuracy in SMS spam detection.

As shown above, many researchers have investigated the same topic using different approaches. Some have focused on security within mobile architecture, including message transfer processes, while others have concentrated on using Machine Learning (ML) models to combat the issue of spam. In general, these approaches are valuable to this project and serve as its inspiration at the point that many techniques related to these approaches have approaches have been implemented in this project.

However, what sets this project apart is its pratical approach involving specific society. Rather than solely relying on existing datasets from platforms like *Kaggle and UC Machine Learning Repository*, this project has actively engaged with people to collect data. It has also integrated some data from these platform datasets to enhance the quality of information.

Furthermore, this project harnesses the latest advancements in machine learning. It utilizes Ensemble Methods to achieve high levels of accuracy. Addionnally, it employs technical parameter tuning, including *GridSearcher and VotingClassifier*. In fact, Grid-

[1] **UCI ML** : The UCI Machine Learning Repository is a popular collection of datasets maintained by the University of California, Irvine (UCI). It serves as a valuable resource for researchers and practitioners in the field of machine learning and data mining. `https://archive.ics.uci.edu/`

[2] **Kaggle :** a platform for data science competitions and datasets. `https://www.kaggle.com/`

Searcher assists in identifying the most suitable parameters required for algorithm models.

Moreover, this project doesn't stop at model development, it extends to the deployment of the models generated through the processes. It provides backend *APIs* for certain platforms interested in learning from these results. Additionally, it outlines the structure of GSM deployment, encompassing the SMSCenter's role in the mobile messaging system.

2.3 Thinking in Machine Learning

Machine Learning comprises a collection of methods and techniques designed to autonomously identify patterns within data. Subsequently, these discovered patterns are leveraged for tasks such as predicting future data points or facilitating decision-taking in uncertain scenarios.

In contrast to conventional programming, where applications are crafted with algorithms tailored to solve particular problems or accomplish specific tasks, ML algorithms take a different approach. They build models that possess the ability to learn independently from input data, accumulate experience over time, and actively contribute to the decision-making process of computers [2].

Nowadays ML algorithms are utilized in many domains of life. In automotive industry for example, they are integrated in drive-systems to treat data got from sensors, cameras for making autonomous vehicles. In huge companies like Amazon they are used in demand-forecasting by ensuring that popular items are well-stocked, reducing the chances of running out of stock or overstocking; they are also used to suggest products that customers are likely to be interested in [20]. Besides, they are used in healthcare systems where they play a significant role since they are able to identify patterns in medical images to detect diseases like cancer; even tailoring treatment plans to individual patients based on their medical history, genetic makeup as well as skin features.

Furthermore, reaching the stage where the ML model is ready to make predictions involves a structured sequence of steps that should be followed in every ML project [12]:

Firstly, the definition of the object and specification : which includes specifying the problem the model aims to solve or the task it should perform. At this stage, all details found serve as project's guiding principles.

Secondly, the preparation and exploration : At this step, ML designer team collects, prepares, examines the data that will be used to train and test the machine learning model. Indeed, the data preparation process encompasses tasks like cleaning, restructuring, formatting, while data exploration involves using visualization techniques to gain valuable insights from the data that is helpful in outlier detection, trend and clusters identification [38].

Thirdly, the model building : This step leads to the construction of the model by selecting the most suitable algorithms and techniques for the specific task. Actually, the model ought to be built by using the prepared data.

Fourthly, the implementation : Once the model is constructed, it needs to be seamlessly integrated into the application or system where it will be put into use. This step involves writing the necessary code and software components to make the model operational, capable of predictions, or generating insights.

Fifthly, the testing : At this stage, the model's performance is assessed using a separate dataset that was not used during training. It is important to ensure that the

model can consistently make accurate predictions and to identify and address any potentials issues of shortcomings.

Finally, the deployment : Once the model is tested, then it is deployed in production environment so that it may be accessible to users or systems where it can provide predictions or insights based on new data.

2.4 Tasks of ML

2.5 Tools and Techniques

In this section we are going to dive into the technologies such softwares and hardwares used to render this project achievable.

Bibliography

[1] Iosif Androulidakis, Vasileios Vlachos, and Alexandros Papanikolaou. Fimess: filtering mobile external sms spam. In *Proceedings of the 6th Balkan Conference in Informatics*, pages 221–227, 2013.

[2] Khalid Anwar, Jamshed Siddiqui, and Shahab Saquib Sohail. Machine learning techniques for book recommendation: an overview. In *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India*, 2019.

[3] Paul Kimumwe Lillian Nalwoga Juliet Nanfuka Edrine Wanyama Wairagala Wakabi PhD Ashnah Kalemera, Victor Kapiyo. State of internet freedom democratic republic of the congo 2019. *Mapping Trends in Government Internet Controls, 1999-2019*, 2020.

[4] Allan Bluman. *Elementary statistics: A step by step approach*. McGraw-Hill Education, 2017.

[5] Jeff Brown, Bill Shipman, and Ron Vetter. Sms: The short message service. *Computer*, 40(12):106–110, 2007.

[6] Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

[7] Guangquan Chen, Weijun Wang, and Xuan Zhou. A survey on sms spam filtering techniques. *Journal of Network and Computer Applications*, 80:149–159, 2017.

[8] Catalin Cimpanu. Simjacker vulnerability exploited for surveillance by at least one nation-state. *ZDNet*, 2019.

[9] Michael Crawford, Taghi M Khoshgoftaar, Joseph D Prusa, Aaron N Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1):1–24, 2015.

[10] John W Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2014.

[11] Prof. Dantu's CSE. Cellular network basics. In *ADV. REAL-WORLD NETWORKS*, volume 14-760. University of North Texas, 2018.

[12] Julian David. Designing machine learning systems with python. *Community Experience Distilled*, pages 381–386, April 2016.

[13] MONUSCO DRC. Protect, stabilize, consolidate peace, monusco's report. *United Nations Organisation Stabilization Mission in the Democratic Republic of Congo*, pages 1–4, 2015.

[14] Cori Faklaris and Sara Anne Hook. Oh, snap! the state of electronic discovery amid the rise of snapchat, whatsapp, kik, and other mobile messaging apps. 2016.

[15] Antonio Ghezzi, Marcelo Nogueira Cortimiglia, and Alejandro Germán Frank. Strategy and business model design in dynamic telecommunications industries: A study on italian mobile network operators. *Technological Forecasting and Social Change*, 90:346–354, 2015.

[16] Suparna Das Gupta, Soumyabrata Saha, and Suman Kumar Das. Sms spam detection using machine learning. In *Journal of Physics: Conference Series*, volume 1797, page 012017. IOP Publishing, 2021.

[17] Kennedy Ochilo Hadullo and DM Getuno. Machine learning software architecture and model workflow. a case of django rest framework. *American Journal of Applied Sciences*, 18(1):152–164, 2021.

[18] Shuang Hao, Nadeem Ahmed Syed, Nick Feamster, Alexander G Gray, and Sven Krasser. Detecting spammers with snare: Spatio-temporal network-level automatic reputation engine. In *USENIX security symposium*, volume 9, 2009.

[19] Marko Hassinen and Smile Markovski. Secure sms messaging using quasigroup encryption and java sms api. *SPLST*, 3:187, 2003.

[20] Ralf Herbrich. Machine learning at amazon. *WSDM*, 535, 2017.

[21] Sunil Kumar Jangir, Manoj Kumar Sharma, and Pawan Kumar Gupta. Design and implementation of sms gateway api for mobile communication networks. *International Journal of Computer Applications*, 151(9):1–5, 2016.

[22] Thomas R Karl, Claude N Williams Jr, Pamela J Young, and Wayne M Wendland. A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the united states. *Journal of Applied Meteorology and Climatology*, 25(2):145–160, 1986.

[23] Veena K Katankar and VM Thakare. Short message service using sms gateway. *International Journal on Computer Science and Engineering*, 2(04):1487–1491, 2010.

[24] M Lavanya and KR Aruna. Sms spam detection using deep learning. *Journal homepage: www. ijrpr. com ISSN*, 2582:7421.

[25] Gwenael Le Bodic. *Mobile messaging technologies and services: SMS, EMS and MMS*. John Wiley & Sons, 2005.

[26] Gwenaël Le Bodic and Hatim Zaghloul. *Mobile Messaging Technologies and Services: SMS, EMS, and MMS*. ABC Publishing, 2022.

[27] Matti Leppäniemi and Heikki Karjaluoto. Mobile marketing: From marketing strategy to mobile marketing campaign implementation. *International Journal of Mobile Marketing*, 3(1), 2008.

[28] A Medani, Abdullah Gani, O Zakaria, AA Zaidan, and BB Zaidan. Review of mobile short message service security issues and techniques towards the solution. *Scientific Research and Essays*, 6(6):1147–1165, 2011.

[29] Ajay R Mishra. *Advanced cellular network planning and optimisation: 2G/2.5 G/3G... evolution to 4G*. John Wiley & Sons, 2007.

[30] Pavas Navaney, Gaurav Dubey, and Ajay Rana. Sms spam filtering using supervised machine learning algorithms. In *2018 8th international conference on cloud computing, data science & engineering (confluence)*, pages 43–48. IEEE, 2018.

[31] Statista's own team of researchers and analysts. *Number of mobile messages worldwide from 2019 to 2023 (in trillions)*. https://fr.statista.com/, 2020.

[32] Chris Park. *A Dictionary of Environment and Conservation*. Oxford University Press, 1 edition, 2012. Online Publication.

[33] Sebastian Raschka and Vahid Mirjalili. Python machine learning: Machine learning and deep learning with python. *Scikit-Learn, and TensorFlow. Second edition ed*, 3, 2017.

[34] Muhammad Abdulhamid Shafi'I, Muhammad Shafie Abd Latiff, Haruna Chiroma, Oluwafemi Osho, Gaddafi Abdul-Salaam, Adamu I Abubakar, and Tutut Herawan. A review on mobile sms spam filtering techniques. *IEEE Access*, 5:15650–15666, 2017.

[35] Houshmand Shirani-Mehr. Sms spam detection using machine learning approach. *unpublished) http://cs229. stanford. edu/proj2013/Shir aniMeh r-SMSSpamDetectionUsingMachineLearningApproach. pdf*, 2013.

[36] Alex Smola and S.V.N. Vishwanathan. *Introduction to Machine Learning*. Cambridge University Press, Cambridge, UK, 2010.

[37] Siyuan Tang, Xianghang Mi, Ying Li, XiaoFeng Wang, and Kai Chen. Clues in tweets: Twitter-guided discovery and analysis of sms spam. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2751–2764, 2022.

[38] Antony Unwin. Why is data visualization important? what is important in data visualization? *Harvard Data Science Review*, 2(1):1, 2020.

[39] Pacifique Zikomangane. A free wireless network in the drc: An answer to internet shutdowns and exorbitant access costs. 2018.