Chibuikem Divine Offor

SPAM IDENTIFICATION IN CLOUD COMPUTING BASED ON TEXT FILTERING SUSTEM

Technology 2025

VAASAN AMMATTIKORKEAKOULU
UNIVERSITY OF APPLIED SCIENCES
Cloud-Based Information Technology

ABSTRACT.

Author	Chbuikem Divine Offor
Title	Spam identification in cloud computing based on text filtering method.
Year	2025
Language	English
Pages	42
Name of supervisor	Johan Dams

This research displays the analysis, development, and training of a software tool designed to block or filter spam messages, preventing them from invading people privacy or stealing sensitive data. This tool is designed to help and protect firms and individual alike by flagging messages with the features and characteristics of spam.

For this to be achieved the use of ML and NLP methods were used, and it was very helpful. Previously other traditional methods were used but due to the advancement of technology attackers found new methods to use. By using tools like ML and NLP, it was able to achieve the goal of this research (Filtering spams and spam identification).

While developing the software, scalability, cost, and reliable cloud platform were all considered. The main objective of this thesis, which is developing a system that have the capacity to identify spams and filter them were achieved.

Keywords Spam Detection, Text Filtering System, Cloud, Natural Language Processing.

CONTENT.

1. INTRODUCTION.

The adopting of cloud computing has in a lot of ways help how we store, use and process data or information. In addition it also came with the ability for it to contain and sustain more traffic inflow in situations where user number increases. When talking about cloud computing as a consumer, they are various sections you can fall into. Either as one who uses the cloud based on needs or necessity, or you make use of shared resources as a service, or as those that pay for the exact services they make utilize. Cloud computing, is really doing a lot of help to businesses in their growth and expansion. Now businesses or companies do not need to build their own data centers, or employ engineers that will fix any problem that arises at their data centers. The birth of cloud computing has effectively scrapped out that and introduce a more cost effective and fast and reliable service for companies. With all these achievements been made, cloud computing still faces some existing threat from elements like fake emails, spamming, etc. Just like when big and mega companies like Amazon and googles faced some security vulnerabilities in 2009 (Bassi & Chaudhary, 2015). When malwares distributions affect user data, it exposes it to threat and therefore the user privacy is truncated.

The purpose of the study is to analyze and study how advanced machine learning techniques and scalable cloud-based solutions are used to tackle the problem of spam. Will be using tools like Natural Language Processing (NLP) and classification models to detect and block unwanted messages, while review spam detection focuses on identifying fake reviews through textual patterns and user behavior. Will be using Jumia as a case study.

1.1 Background and Motivation.

As we all know we are in a digital world where most of our daily financial transactions are digitalized, which is giving birth to new and more dynamic digital platforms to enhance and make it easier for people to transact easily and seamlessly. The digital space has proven to reduce the work load for business people and companies, also in terms of reach, the digital space has helped a lot of businesses expand its reach within its geographical location.

Nigeria as it stands now is the fourth largest economy in Africa with about $180bn in possession with a population of over 150 million inhabitants, which position it as a very attractive opportunities for businesses to leverage on.

In recent times, the rise of internet users in Nigeria has significantly increased compared to previous years. From the year 2000-2008, up until 2016 there has been an increment of internet users totaling a number of 90% (Ibam, et al., 2018). With this development it gave businesses the ability to leverage the internet space to operate and connect widely to their potential customers.

This has given birth to the creation of many ecommerce platforms like Jumia, which will be used as a case study. Jumia as one of the popular and biggest ecommerce platforms in Africa and in Nigeria, is seeing as the first when it comes to online retailing and sales.

Jumia connects thousands or more online vendors who offer services or have products to sell to available buyers, they achieve this while operating on a cloudbased infrastructure where scalability, cost minimization, and efficiency is prioritized. While this achievement is been made, there are still factors that can affect the infrastructure or pose a threat to it. Take for example spam which affects multiple facets of the

infrastructure. It also stretches out to fake product reviews and product listing not to mention but a few.

Due to the modus operandi of Jumia where most of their communication with the vendors are primarily with the use of email, from one end to another. This tends to pose a risk to users of this platform. Spammers use this avenue to subtle send in spam messages searching for an avenue to strike. When all these are not carefully checked it pose as a threat to the reputation of Jumia and therefore, might result to a decline in the way people use the e-commerce platform. In a situation where mails sent to users pretend to look like that of the original Jumia mail, it leads to loss in financial asset and data.

When talking about Phishing attack we have several types of Phishing attack;

- **Spear Phishing**: In this scenario, the attackers usually customize emails to look like a particular institution credential, mostly using private information of the organization to appear credible.
- **Deceptive Phishing**: In this situation, attackers tend to impersonate companies, then trying to gain access to their personal details or passwords. They further black mail the user to do the bidding of the hackers.
- **Link Manipulation**: The attackers send a malicious or altered link where by the link re-directs you to a different website whereas, it is supposed to take you to the original website mentioned in the link.

There are many more types of Phishing attacks out there, users are advised to always look out for situations like this and be mindful and also techniques on how to identify these attacks. Most of phishing links contains malicious code. Phishing as we all know is a type of social engineering and it is mostly used for hacking of email. The hacker might

send information that are false which will cause the user to lose sensible data (Bhavsar et al., 2018).

Driven by the motivation to systematically address these challenges, in this research I will be exploring and categorically stating how advance machine learning technique and scalable cloud solutions can be used to combat spam on Jumia.

Currently, NLP Technology is been utilized to detect phishing and malicious mails (Salloum et al., 2021). Machine learning models like deep neural network and classification algorithms, possess the capacity to point out patterns that are potential spams; while doing this it also has the capacity to scale and mitigate properly.

This research further talks on how Jumia tackles spam by leveraging on existing tools to safe guard its e-commerce space, make it a safe place for their users to transact and do business easily.

1.2 Statement of problem and research problem.

Because of how vast and complex the cloud environment is, it has become difficult to tackle spamming using the traditional way thus making way for the development of more advanced tools and techniques. In an ecommerce environment like Jumia, where a lot of user data are used to process payments, communication and product listing, makes it so challenging to use the traditional method for detecting spam.

This research will focus on ways to tackle spam in cloud-based systems text filtering techniques. By doing this, I will be using existing tools like NLP and ML.

Below are some of the research questions;

1. What techniques are most useful for spotting spam in Jumia’s cloud environment?

2. How to apply machine learning and NLP to detect fake emails and fraudulent activities specific to Jumia's ecosystem?

3. What strategies can ensure spam detection systems scalability and realtime efficiency in Jumia's cloud infrastructure?

1.3 Objectives of study.

The main goal of this study is to put into considerations of Jumia's challenges in fighting or avoiding spam within its cloud-based systems, by developing text filtering systems that have proven to be accurate. Study materials that have been used previously by other researchers will be visited, beaming our focus on text filtering and ML. This will help in creating a foundation for Jumia as an e-commerce platform on how to use cloud technologies to avoid spams.

The study will also use available data to know the type of spams that are common on Jumia's platform.

Thirdly, to design and develop a text filtering system that uses machine learning (ML) and Natural Language Processing (NLP) techniques to detect spam. It will prioritize scaling and efficiency and cost.

2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK.

This chapter will be written based on the considerations of existing papers and research work about spam and text filtering system using ML and NLP techniques. The theoretical frame work will be an in-road of existing theories that added to the research of spam and text filtering systems using ML and NLP techniques.

2.1 Cloud computing: Background and historic trends.

Previously, before cloud computing and its services came into play, individual businesses and corporations have data centers across different location or in their business premises, this made it expensive for companies to manage and also increased the work force in companies because they will need the services of those who will manage these centers for them and also, most times these companies suffer downtime because of either lack of storage or cost. This gave birth to the coming of cloud computing and its services. It helped businesses solved the problem of cost, scaling and data storage without the fear of losing them.

Some years back, cloud computing have been gaining momentum in the technology sector globally. This concept has been really helpful to businesses especially small scaled business, because you pay for the services that you use on the internet. When businesses make use of some certain services on the cloud, they pay for the exact service which they are using. Businesses are not charged outside of what they did not use, a lot of businesses have considered this very helpful. In a situation where a

company makes use of a cloud service provider in storing data, that company is charged for the storage of data nothing more, nothing less a lot of businesses have considered this a win-win situation (Bass & Chaudhary, 2015).

Before the coming of cloud computing, most companies find it very challenging and difficult to manage data centers or buy servers. When talking about cloud computing as it stands today, we have three major company providing cloud computing services namely;

- Amazon web services (AWS).
- Google cloud providers (GCP).
- Microsoft Azure.

The coming of these cloud services eased a lot of stress on a lot of companies world-wide, making operations easy and scalable without delays at the same time been cost efficient. Companies no longer have to worry about data centers; they now pay for a subscription fee on a cloud provider service and use of whatever services or assistance they need.

It helps them to scale up easily and scale down easily.

Furthermore, cloud computing is widely known as a direct source of a strong software that performs lots of given task easily and accurately with minimal cost (Foster & Gannon, 2017).

The swift growth and the wave of cloud computing makes it an interesting area in IT for research. In 2006, big cloud service providers like google, amazon etc. Introduced new cloud computing system to the market (Yang & Tate, 2012). The term or word cloud computing, was newly initiated to the IT world which refer to the computing consumption where users pay according to what they use (Armbrust et al., 2009).

Furthermore, cloud computing is comprises of five (5) characteristics as namely;

- On-demand self service
- Broad network access
- Resource poling • Rapid elasticity
- Measured services.

As stated by the United States National Institute of Standards and Technology (NIST), (Mell & Grance, 2011). Cloud computing has on a very large scale helped a lot of businesses ranging from agriculture, academics, aviation, and commerce. Its flexibility gives it an edge in the IT world and it aids business to scale up and scale down easily.

2.2 Analysis of cloud service providers (CSP's).

Cloud service providers (CSPs) in recent times have in the modusoperandi of many businesses. It now helps business manage financial resources and also helps businesses in storage of data. The common and popular Cloud service providers that are rendering lots of services to business are;

- Amazon web services
- Google cloud platform
- IBM cloud
- Alibaba cloud.

These companies offer different types of services ranging from

Infrastructure as a service (Iaas), Software as a service (Saas) to Platform as a service (Paas). This makes it very easy for companies to store and manage data on the cloud.

Amazon web services have the largest market share compared to others like google cloud and Azure. In an article written by (Richter, 2025), It was discovered that in the Q4 of 2024 Amazon market share accounted for about 30% of the global market for cloud infrastructure, while Microsoft Azure followed with 21% and Google cloud at 12% (RENO, 2025).

Table 1; Q4 2024 Market share of cloud service providers by Synergy energy group, (Synergy Research Group, 2025).

Cloud Providers	**Market share.**
Amazon web services	30%
Micro soft Azure	21%
Google cloud platform	12%
Alibaba cloud	4%
Oracle	3%
Sales force	2%
IBM cloud	2%
Tencent Cloud	2%

In addition, Aws offers large amount of services which makes it stand out from others. Services like Elastic Cloud Compute (EC2), Simple Storage Service(S3) and many more. It also has an advantage of the vast and massive data centers littered in several locations around the world.

Figure 1. Amazon web services growth since 2004 by (Mufti et al., 2021)

Microsoft Azure was introduced to the public in 2010; it generally works perfectly with existing Microsoft software.

Google cloud platform (GCP) leverages google infrastructure to offer several tools like machine learning, computing etc. It is a very good option for data driven application.

Figure 2. Google cloud platform (GCP) since 2011 by (Mufti et al., 2021)

Each providers have its own unique strength like the ability to scale, cost efficient, etc.

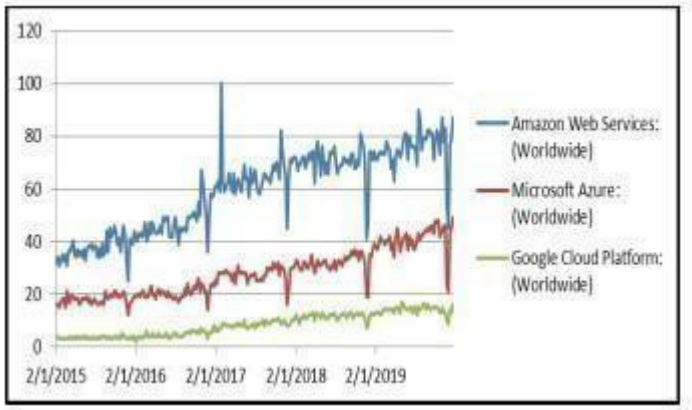


Figure 3**.** Amazon web services (AWS) vs Micro soft azure vs Google cloud platform (GCP) by (Mufti et al., 2021)

2.3 Machine learning techniques in spam detection.

Considering the point discussed above, and also considering the fact that we are in a generation where the use of internet is on the rise, that means most of our communication are done over the internet. Communicating over the internet is quite easy and fast in my own opinion, but the challenge and what sometimes maybe considered as a digital nuisance most of the times are spam messages.

Imagine waiting on an email message from a co-worker or a staff, or from your procurement team, and whenever an email pops on your device you check with the hopes that it is an important message you have been waiting for, no one likes that, it not only serves as a time waster, also most of the spam related messages have malicious content hidden in them (Blanzeiri & Bryl, 2008).

So therefore, spam are unwanted messages or sometimes email that is sent out by a spammer or an attacker with the intention to steal or cause damage to people devices. This message can be sent either by email or any other medium (Alghoul, et, al., 2018). With these problems highlighted, there is need for systems to be developed that will check and

protect users from such. The word “Spam” is gotten from the Monty Python Sketch (Petersen, 2018).

Initially companies use the traditional way to check and avoid spam from harming or causing damage to their systems. Methods like static and rulebased approach worked well. Before spammers became more intelligent and sophisticated, keyword screening worked well but with time and advancement in technology it wasn’t very useful again.

In the rule-based approach some certain words were deemed as red-flag. Words like “Money”, “Free”, “Win” etc. whenever such words appear in an email, it is flagged as spam. To bypass the systems spammers or attackers devised a new method to write these words. Instead of the word “Win”, they write “W1N”, Instead of “Money”, they write “M0ney”.

The traditional way (Rule-based approach) couldn’t detect these patterns.

Machine Learning (Alpaydin, 2020), is one of the crucial and important application of artificial intelligent (AI). It enables the system to learn and upgrade its functionality without programming (Alpaydin, 2020). The main use of machine learning algorithm is mainly to develop tools that are automated, to have access to data and make use of the data for trainings. The first phase is to label your data, which is also known as training datasets (Ahmed et al., 2022). This can be real life experiences, reviews, examples or feedback, this will enable it recognize patterns to it will be able to make correct and accurate decisions in the future (Ahmed et al., 2022).

Machine Learning primary objective is to learn and study patterns without the assistance of any human (Ahmed et al., 2022). Because of how vast and dynamic Machine learning is, it has the capability to handle large sum of data and most of the result are hardly wrong and Machine Learning will be suitable for filtering spam because of the advancement in technology

been used by spammers (Ahmed et al., 2022). Some of the advantages is that it performs very well with accurate datasets and the disadvantages is that when some letters are replaced with characters it might not be able to detect it e.g. “0ffer” in place of “Offer”.

Below is a diagram by (Ahmed et al., 2022). Displaying the types of Machine Learning methods.

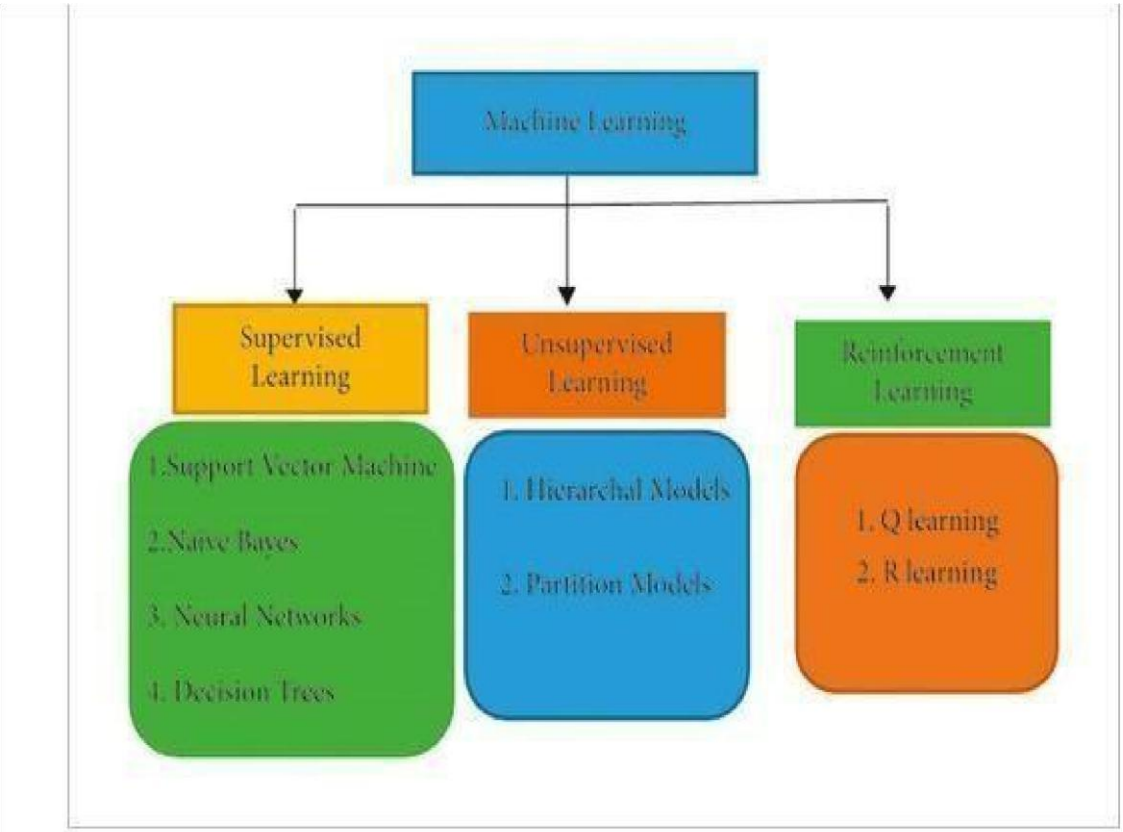


Figure 4 Types of Machine Learning methods by (Ahmed et al., 2022).

From the diagram above, we will discuss more on SVM, Naives Bayes,

Neural Network, which are under supervised Learning. Supervised Learning (Kotsiantis, et, al., 2007) methods are methods where data needs to be labeled.

At the initial stage, labeled training data are made available to these models for extensive trainings, which enable the models to perfectly predict future event (Ahmed et al., 2022).

Supervised Machine Learning (SML) uses enough labeled data for trainings, it helps it to predict the incoming or new data based on the training in essence, this type of learning can be used to tackle spam problem (Ahmed et al., 2022).

Below is a diagram by (Ahmed et al., 2022) on how supervised machine learning operate.

Please! note; when training data endeavor to use accurate data this will help (SML) in training accurately and giving you good and perfect result.

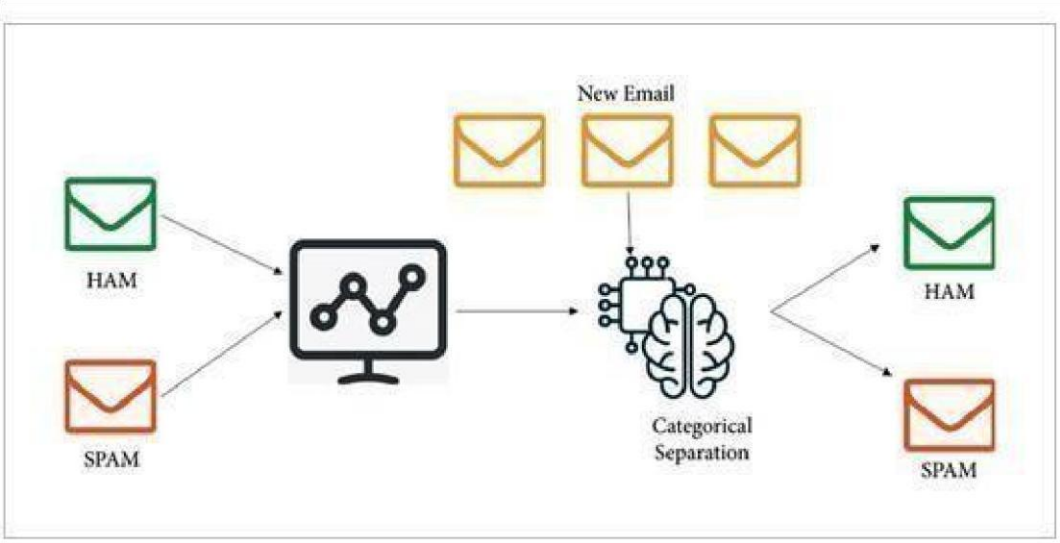


Figure 5 Supervised Machine Learning Diagram by (Ahmed et al., 2022)

Jumia, is one of the biggest e-commerce platforms in Nigeria, and a leading ecommerce platform in Africa primarily base its communication with its users via email, making use of Supervised Machine Learning method will be very helpful to the reputation of the company because it will not only help tackle potential spam messages, it will also protect user-data.

2.4 Role of natural language processing in spam detection.

Phishing is a type of cybercrime where very sensitive information of users are hacked into and made use of without the consent of the original owner (Ora, 2020). A researcher (Li, et, al., 2019) came up with a remedy by implementing

Word2Vec as an extraction technique that gathers the features from HTML codes.

To enhance performance, an integrated model was developed utilizing feature extraction techniques alongside algorithms such as LightGBM, XGBoost, and

Gradient Boosting Trees (Ora, 2020), (Li, et al., 2019). This combination led to improved outcomes in spam detection tasks. Additionally, Liew et al. (2019) introduced a real-time phishing tweet detection mechanism.

Their approach devised a supervised machine learning model, specifically Random Forest, leveraging eleven carefully selected features to identify phishing URLs embedded in tweets. This system achieved a high accuracy rate of 94.5%.

Another researcher (Sahingoz et al., 2019), discovers a phishing website by analyzing, extracting and examining from the URL, the random word detection module is applied which goes on to break down URL into micro sections, then the micro sections will be used to determine if the website of genuine. Several machine language (ML) methos were used like SVM, Naives Bayes etc.

In the aspect of Jumia, where spam appear in different forms, it is important that NLP methods like Word2Vec should be instrumental.

2.5 challenges in spam detection.

An inquiry of spam SMS filtering was carried out on the UCI machine learning dataset in 2015 by (Kim, et al., 2015) where she used the feature selection technique like the Naives Bayes etc. After the inquiry, it was observed that Naives came out with high accuracy of 94%.

Several other research were conducted and it was observed that SVM and Naives Bayes performed very well (Ora, 2020). Even with the complexity of Spam filtering and detection, there are still some challenges that affect spam detection.

Earlier when the traditional method was used to avoid spams, more sophisticated method was adapted by attackers, therefore, the using of ML and NLP to combat spamming is a great idea but there must be avenue to advance in more research and technicality so as to counter any future technique that will be adopted by the attackers.

3 RESEARCH DESIGN AND METODOLOGY.

To find practical spam detection solutions, the study uses a hybrid methodology that combines exploratory and experimental techniques. To comprehend the difficulties presented by spam in cloud settings, especially in the Nigerian context, exploratory research was carried out. Case studies were used to acquire information about the system needs and performance expectations, such as how spam affected Jumia Nigeria's operations. The spam detection system was designed and implemented through experimental research, which iteratively tested its performance under various scenarios to improve its scalability and accuracy.

The research methodology is structured as follows:

1. **System Development**: Build a machine learning-based spam detection system that can operate within a cloud computing environment.
2. **Data Collection**: Gather relevant datasets from cloud-based platforms to train and test the system.
3. **Algorithm Selection**: Implement and compare different types of machine learning algorithms to choose the best-fit model for detecting spam.
4. **Evaluation**: Use performance metrics to check the system’s accuracy, precision, recall, and computational efficiency in a cloud-based environment.

3.1 Data collection

The goal of data collecting was to compile a representative dataset of both spam and lawful communications. This dataset was derived from anonymized communication records supplied by Jumia Nigeria using

publicly accessible spam corpora, including the Enron Email Dataset. Data collection was conducted with ethical concerns in mind, guaranteeing user privacy and adherence to Nigerian data protection laws. For efficient model training, the gathered data was pre-processed to eliminate duplicates, standardize text content, and balance class distribution.

The quantity and quality of data have a major effect on the spam detection system's achievement. The following methods of gathering data will be used for this study:

Platform Data: The primary dataset will be gathered from cloud-based communication platforms such as emails, social media messages, and customer reviews from platforms like Jumia Nigeria. The data will be collected with permission and anonymized to protect privacy.

Spam Datasets: Spam datasets that are publicly available such as **Enron Spam Dataset** and **LingSpam Dataset**, are going to be used to train initial models and evaluate performance.

Synthetic Data: To augment the dataset, synthetic spam data will be generated using rule-based systems or other spam content generators to simulate real-world spam scenarios.

3.2 Data pre-processing.

Text Preprocessing: The collected data will undergo series of NLP preprocessing which include tokenization, stemming, lemmatization, and put a stop to word removal, to ensure that the input data is clean and standardized for analysis.

Feature Extraction: Key features such as word frequency, message metadata (sender, timestamp), and semantic content will be pulled from

text using **TF-IDF (Term Frequency-Inverse Document Frequency)** and **word embeddings.**

Data Labeling: A manual labeling process will be employed to arrange data as either "spam" or it is "legitimate." This labeled dataset will be used to train supervised learning algorithms.

3.3 System development.

The proposed spam detection system will be developed using machine learning algorithms and NLP techniques. The development process involves the following key steps:

During this phase, the original data was processed and changed into structured formats ready for analysis. Techniques such as tokenization, stop word removal, stemming, and lemmatization were applied to reduce noise in textual data.

1. **Feature Engineering Phase**: Features were extracted from the processed data to enhance model performance. Common features included Term FrequencyInverse Document Frequency (TF-IDF) scores, ngrams, sentiment polarity, and metadata attributes such as sender reputation and message length. NLP techniques were employed to derive semantic and syntactic features from message content, enabling the version to differentiate between spam and legitimate communications based on contextual cues.

2. **Model Training and Deployment Phase**:

   Various machine learning algorithms, even Logistic Regression, Support Vector Machines (SVM), Random Forest, and Neural Networks, were trained and tested using the prepared dataset. Deep learning models such as Long ShortTerm Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) were explored to improve classification accuracy for

unstructured text data. Hyperparameter tuning and cross-validation were performed to optimize model performance.

3.4 Evaluation Methodology.

Key performance indicators like accuracy, precision, recall, and F1 score are used in the evaluation process for spam detection system. While precision assesses the ratio of detected messages that are considered spam, accuracy gauges the system's overall correctness. The system's ability to capture all real spam communications is determined by recall, and the F1 score offers a balanced indicator of both precision and recall. In order to guarantee that the system functions effectively in a variety of scenarios, processing speed, scalability, and real-time evaluation are also taken into account. This makes the system appropriate for real-time deployment in cloud environments.

Accuracy.

Accuracy measures the overall usefulness of the system in differentiating between spam and messages that are non-spam.

- It calculates the proportion of correctly classified messages (both spam and non-spam) out of the total predictions.
- Formula: Accuracy = TP+TN/TP.
- Where TP is True Positives
- TN is True Negatives
- And TP is Total predictions.

Limitation: Can be challenging when data sets are imbalanced.

Precision.

- Precision evaluates the quantity of the messages that identified as spam were actually spam.
- It measures the system's ability to avoid falsely marking legitimate messages as spam (False Positives).
- Formula: Precision = TP/TP+FP • Where TP stands for True Positives.
- And FP stands for False Positives.
-

Importance: High precision means fewer false alarms (legitimate messages mistakenly classified as spam).

Recall.

- Recall measures how well the system catches spam messages.
- It calculates the proportion of spam messages that were correctly classified out of all real spam messages.
- Formula: Recall = TP/TP+FN.
- Where TP = True Positives.
- And FN = False Negative.

F1 Score.

- The F1 Score is the harmonic mean of precision and recall.
- It provides a balanced evaluation, especially when there's an imbalance between precision and recall.
- Formula: F1 Score = 2*PR/P+R
- Where P&R = Precision and Recall.

Importance: A high F1 score indicates a well-balanced spam detection system

Metric.

- This measures how quickly the system processes and classifies messages.
- Faster processing speed is essential for real-time spam detection in cloud environments.
- **Metric Example:** Messages per second (e.g., "The system processes 1,000 messages per second").
- **Trade-off:** A very fast system may sacrifice accuracy for speed.

In this section, we talk about scalability;

- This tests how well the spam detection system performs under increasing workloads.
- The system should handle large volumes of messages efficiently without degrading accuracy or speed.
- **Example:** The system is tested with datasets of different sizes (e.g., 10,000, 100,000, and 1,000,000 messages).
- **Expected Outcome:** Minimal drop in performance as data volume increases.
- This assesses whether the system can detect spam messages as they arrive in real-time.
- A real-time system should have minimal latency and high accuracy.
- **Example:** If a spam detection system is integrated into an email service, it should classify incoming emails instantly without delays.
- **Challenge:** Balancing real-time speed with accuracy and security.
-

3.5 Research validity and Reliability.

During and after research your results must be accurate and reliable, when another researcher tests your research with the data provided by you, will give the same result?

So, the right method needs to be employed when carrying out research.

Your work must be reliable and accurate.

3.6 Limitations of study.

This study has a few challenges that could impact the results. One major constraint of this study is the in ability to source real-world labeled datasets from Nigerian cloudbased companies like Jumia. Without enough high quality data, training the spam detection model effectively could be difficult. Another challenge is the everchanging format of spam. Spammers, always evolve their approach, so while the model might perform well in a controlled test environment, it may struggle to keep up with new types of spam in real-world scenarios. For instance, a platform like Jumia, which relies heavily on customer interactions and email communications, may face different spam threats compared to other cloud-based businesses. The study faces resource constraints, particularly in terms of computing power. Training and deploying

a spam detection system requires significant processing power, and limited resources could affect both the speed and scale at which the model is developed. Despite these limitations, one of the study aim is to provide valuable insights into improving spam detection in cloud environments, with a focus on platforms like Jumia that depend on secure and reliable communication systems.

3.7 Ethical considerations.

When building a spam detection system, it is important to put ethics into considerations. There is data privacy. Since the system will scan messages, we need to make sure user data, including messages from Jumia, stays safe and is not misused. Encryption and other security measures should be in place to protect sensitive information. Another issue is bias. If the system is trained on limited data, it might wrongly classify some messages as spam, causing problems for businesses and users. It is also important to be clear about how the system makes decisions so people can trust it. User consent is another key point, people should have idea of how their private data is being used and have the right to opt out if they want. We need to be careful about filtering messages. The goal is to block spam, not important messages, so the system must be fine-tuned to keep things fair and effective.

3.8 Anti-spam related method.

Spam filtering is useful in shielding users and organizations from dangerous and possible harmful emails.

There are three primary techniques used in spam detection, namely;

- Rule Based Filtering
- Blacklist and Whitelist Filtering
- Statistical Filtering

3.9 How spam identification works.

Spam filtering systems follow a well detailed process to detect and block or flag unwanted emails while ensuring that important messages reach users. The process begins when an email is received by the mail server. Basic

checks are carried out by the systems, like checking if the sender's address is genuine and making sure that the email is subjected to a proper formatting and communication protocols. This helps to filter out incorrectly formatted messages or those from unrecognized sources (potential spams).

The system analyzes the email header and sender details. This is done by checking the IP address, domain name, and authentication records to determine whether the sender is legitimate or considered harmful. Some Methods such as IP reputation checks and domain authentication using DMARC and some other tools to help identify fake or spoofed emails. For example, if an email claims to be from Jumia but fails authentication checks, it may be flagged as suspicious or blocked.

Once the sender is checked and confirmed genuine by the system, the next thing is for the email content to undergo checks. This step involves scanning the message body, attachments, and embedded links for signs of spam. The system looks for keywords commonly used in spam messages, such as "free money" or "urgent action" required. Also, links are checked to see if there are from a known phishing site.

To make spam detection more equipped, blacklists and whitelists are used. A whitelist contains approved senders, ensuring that their emails are always delivered, while a blacklist includes known spam sources that are automatically blocked. These algorithms analyze past patterns to classify messages based on various factors, such

as formatting, sender behavior, and message structure. This approach helps detect new and evolving spam tactics that may not yet be in predefined rules.

After analysis, the email is either delivered to the recipient's inbox, placed in the spam folder, or completely blocked. If a message is flagged as spam but the user marks it as important, the system learns from this feedback

and adjusts its filtering rules accordingly. Similarly, if a phishing email is mistakenly delivered to the inbox and reported by the user, future emails of the same nature are handled more effectively.

Given the increasing reliance on email communication in Nigeria, businesses like Jumia, banks, and government institutions must implement strong spam filtering systems. Without proper filtering, they risk exposure to phishing attacks, malware, and email fraud, which can lead to financial losses and data breaches. A welldesigned spam filtering system ensures smoother communication, enhances security, and helps businesses maintain trust with their customers.

The process concludes with the system ready for the next incoming email, ensuring continuous learning and adaptation to new spam patterns.

3.10 Optimization process of an anti-spam system.

In many organizations like the production sector, health, academics and marine etc. also in the private activities of people, emails have risen to a significant height in terms of popularity, this is as a result of the increase in the use of the internet. So far, email has been proven to be the most affordable and easy to use platform for communication in many organizations. Even at this, it is still faced with some challenges others will consider as disturbing (Olatunji, 2019).

Spam are messages or texts that are seen as a digital nuisance, when you are expecting important messages from a partner or colleague, it happens to be that it is a message from an unknown source, not only that, such messages or email most of the times contains some rudiments of

dangerous links. According to (Ceci, 2024), over 250million emails were sent out in a minute in December 2024 refer to the diagram below.

Figure 6 Media usage in Dec 2024 by (Ceci, 2024).

Due to the rise in credential theft, and individual data been exposed as a result of spam, a lot of tools have been developed to curtail these issues. And the use of ML and NLP has proven to be very effective in detecting spams and getting rid of them.

A researcher (Rusland et al., 2017), carried out research on spam filtering, where he performed the analysis using machine learning. Several datasets

were used based on the evaluation value of accuracy, F-measure, precision and Recall. This research made use of three core steps for filtering the email the steps include; • Processing • Feature selection

• At last.

Spam detection and filtering using ML and NLP is aimed at spotting spams and getting rid of them, also doing no harm to legitimate emails. This is achieved by training a large number of datasets, indicating ‘Spam’ or ‘ham’. The tool study and adapt to the pattern and is able to distinguish between the two. Furthermore, when users mark legitimate email as not spam, the system automatically learns it and adjust accordingly.

3.10.1 Cloud service provider selection and integration.

The choice of a cloud service provider is critical for the deployment of the spam detection system. For this study, the selection process was guided by factors such as scalability, security, reliability, cost-effectiveness, and compatibility with advanced machine learning and natural language processing (NLP) tools. Major cloud service providers, including Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP), were evaluated based on these criteria.

3.10.2 Cloud service provider selection.

AWS was chosen as the preferred provider due to its comprehensive suite of machine learning services, robust security features, and global infrastructure, which ensures high availability and scalability. AWS SageMaker, a fully managed machine learning service, was particularly instrumental in training and deploying the spam detection models. Additionally, AWS's Elastic Compute Cloud (EC2) and Simple Storage

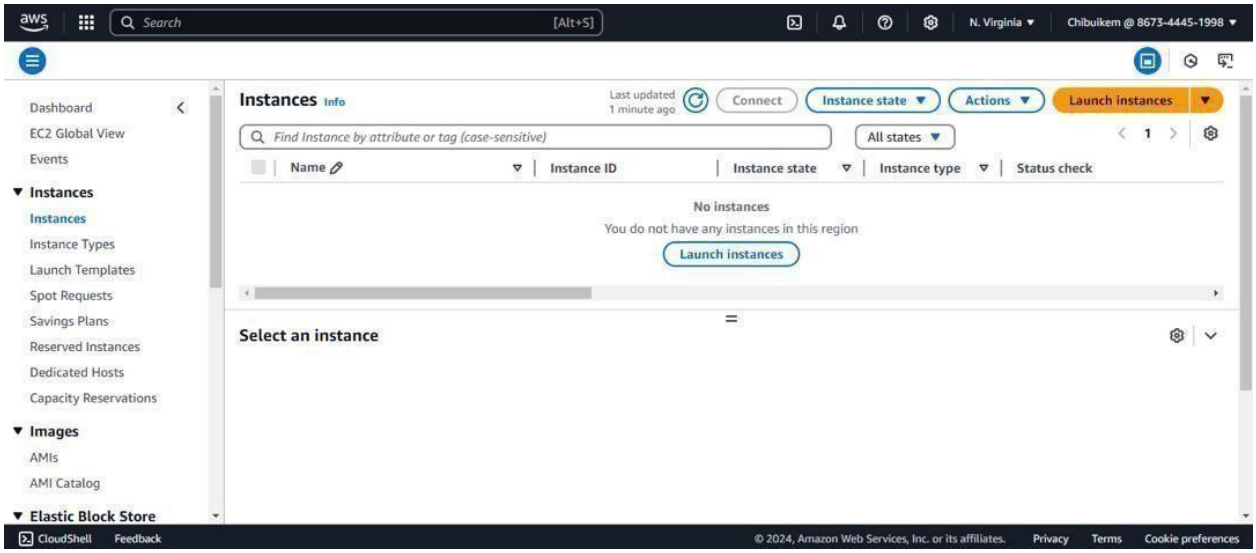
Service (S3) were utilized for computational and storage needs, respectively.

3.10.3 Security considerations.

Given the sensitivity of spam detection and the potential for handling confidential data, AWS's agreement with global standards such as ISO 27001, SOC 2, and GDPR was a determining factor. The platform's encryption features for data at rest and in transit, along with its Identity and Access Management (IAM) policies, ensured a secure environment for the system's operation.

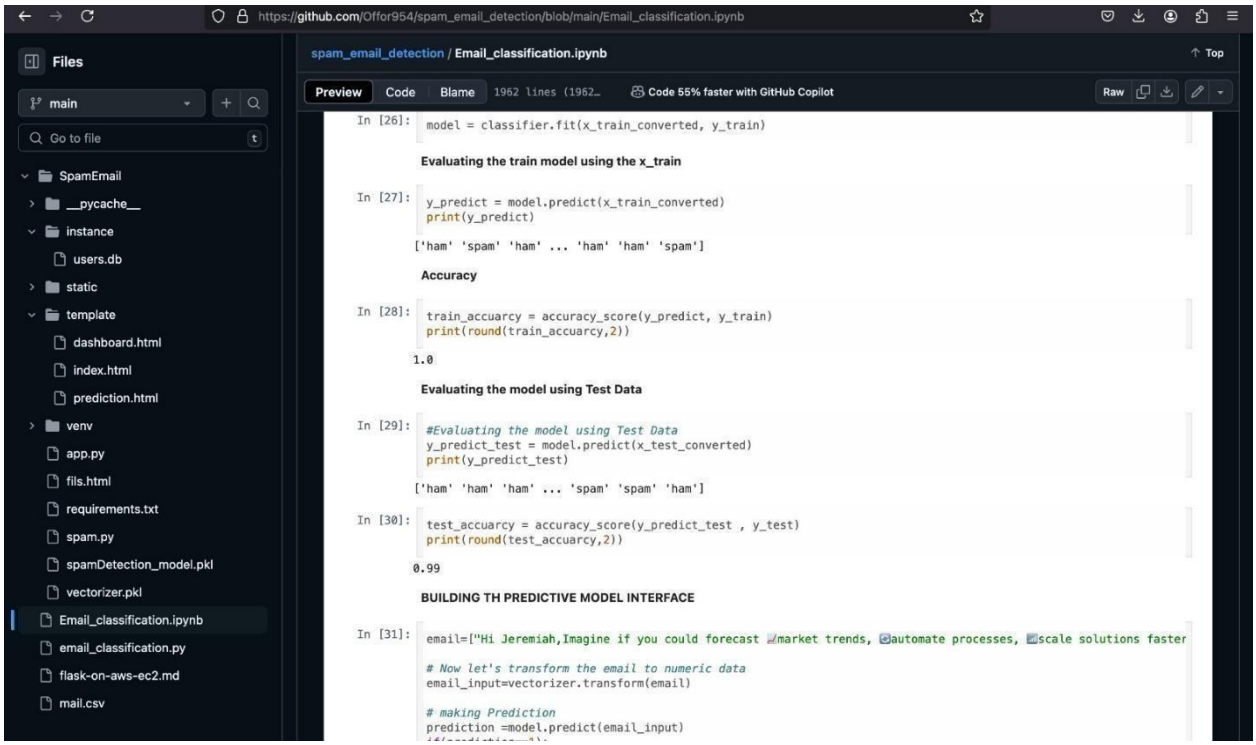
3.10.4 Integration process.

1. **Environment Setup**:



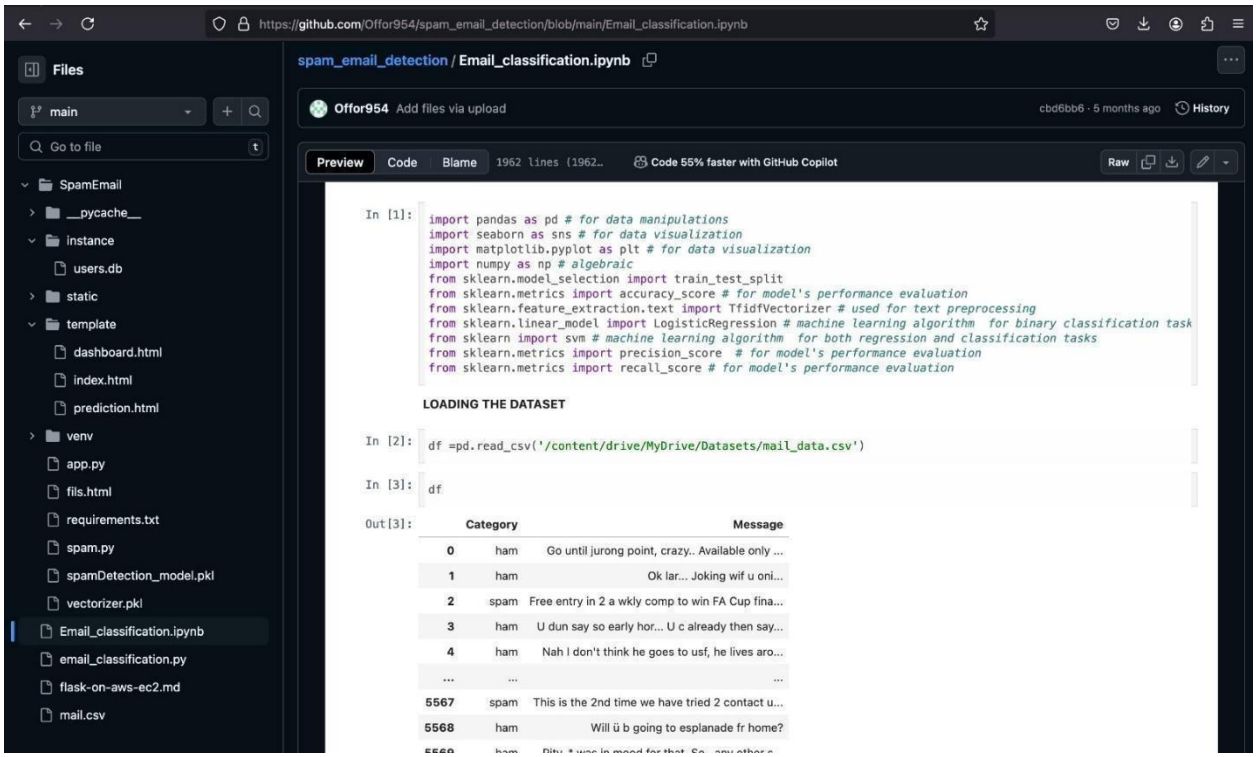
Virtual machines (EC2 instances) were provisioned for model training and real-time inference. Storage buckets (S3) were configured to securely store training datasets, pre-processed data, and model artifacts.

2. **Machine Learning Workflow**:



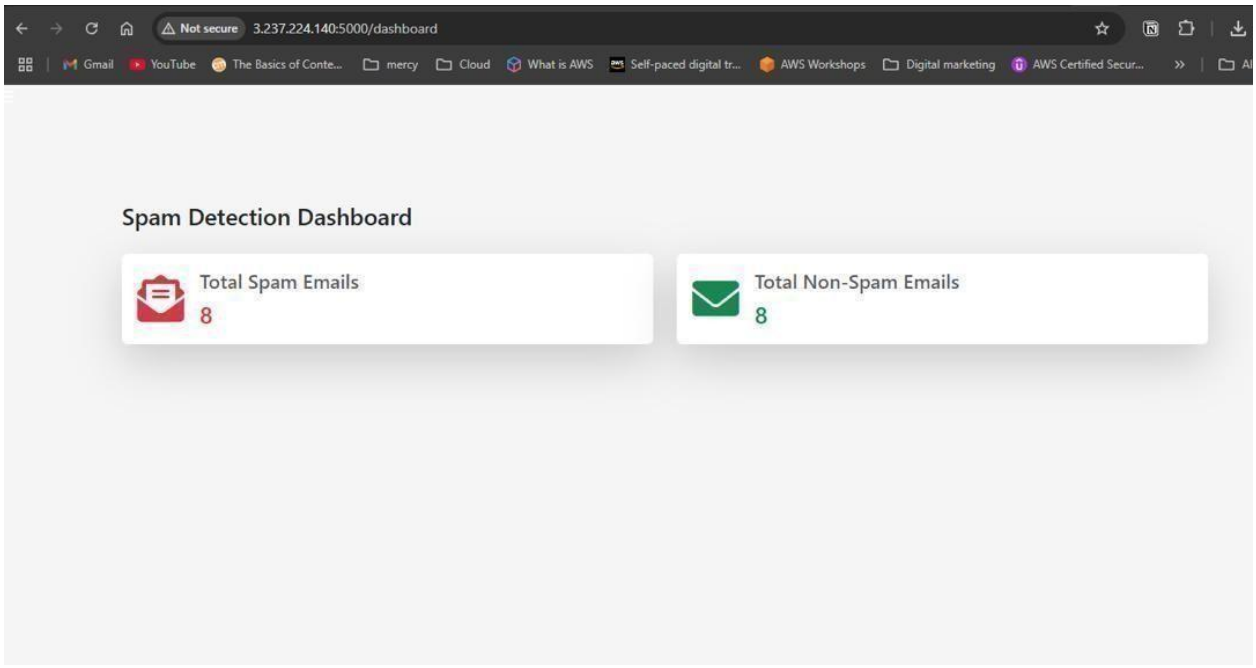
AWS SageMaker facilitated the development of the spam detection models, enabling seamless integration with other AWS services. The models were trained using datasets stored in S3 and fine-tuned on GPU-optimized EC2 instances to accelerate the training process.

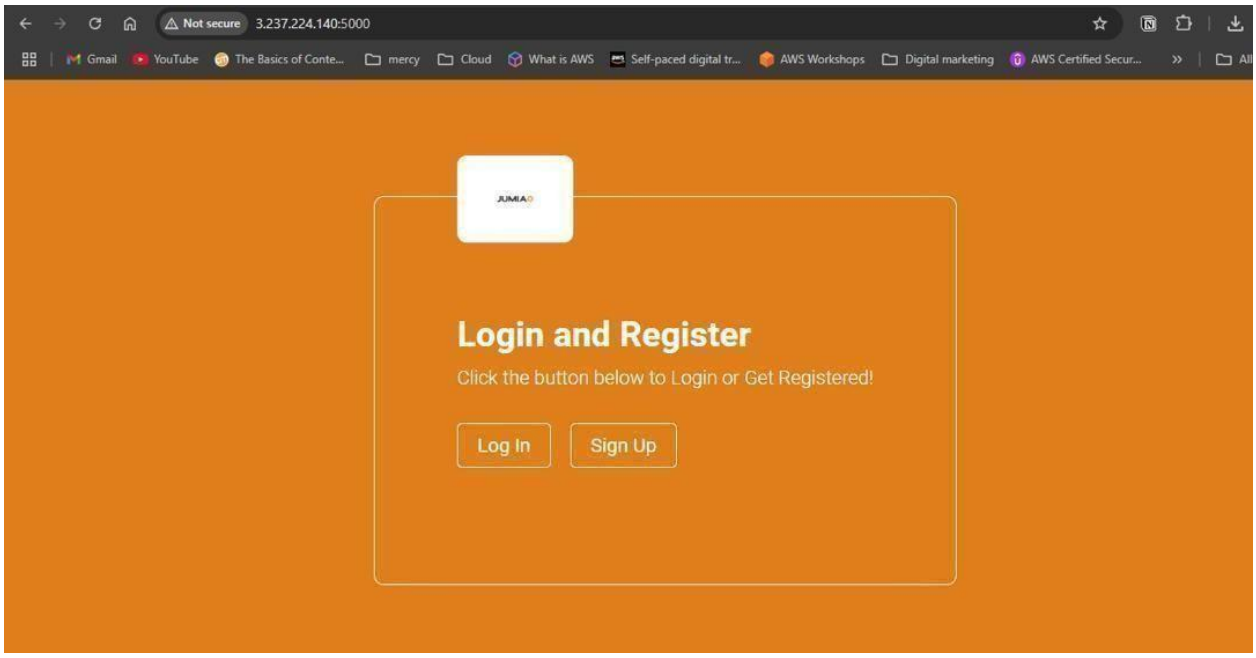
3. **Data Pipeline Configuration**:



A data ingestion pipeline was established to streamline the flow of incoming messages for spam detection. AWS Lambda functions were used to preprocess incoming data in real time, while AWS Glue was employed for data cataloging and ETL (Extract, Transform, Load) operations.

4. **Deployment**:





The trained models were deployed as endpoints using SageMaker, allowing integration with external applications through REST APIs. This setup ensured realtime spam detection while maintaining low latency.

5. **Monitoring and Optimization**:

AWS CloudWatch was utilized to monitor system performance, detect anomalies, and ensure optimal operation. Metrics such as model inference time, request throughput, and detection accuracy were tracked, and adjustments were made as necessary.

3.10.5 Considerations for Nigerian Organizations.

The integration process accounted for specific bottle necks posed by Nigerian organizations, like limited internet bandwidth and in accordance with local data protection regulations. For instance, the system was configured to support hybrid deployment, combining cloud-based inference with on-premises processing where necessary to optimize performance. Additionally, AWS’s Africa (Cape Town) region was selected for data residency, ensuring compliance with regional regulatory requirements

4 RECOMMENDATION FOR FUTURE WORK.

The integration process accounted for specific bottle necks posed by Nigerian organizations, like limited internet bandwidth and in accordance with local data protection regulations. For instance, the system was configured to support hybrid deployment, combining cloud-based inference with on-premises processing where necessary to optimize performance. Additionally, AWS’s Africa (Cape Town) region was selected for data residency, ensuring compliance with regional regulatory requirements

5 CONCLUSION AND SUMMARY.

The introduction of internet to the world, has brought its own challenges from cyberattacks to spamming to corporate bodies suffering immense loss in their business etc. This has also made a lot of government bodies worldwide to draw up laws that will control their digital space, protecting citizens data. The EU also have a regulatory body that checkmates this and they are very serious on preventing the exposure of data.

The invention of a lot of digital technology has helped in combating these issues, one is what we discussed in this thesis using ML and NLP to combat spamming. Spams have caused heavy losses for organizations and individual alike. Before the usage of ML and NLP to combat this challenge, a traditional method was used but, as times goes on the perpetrators became more tactical and smart, this act made researchers to also invent more sophisticated methods to tackle them, from articles and practicals conducted by a lot of researchers it has been proven that this new method which is using ML and NLP to filter and detect spam, has achieved over 80% accuracy.

Furthermore, I will suggest that research should continue in the creation and advancement of technologies that will checkmate this. Because when newer technology is invented to prevent these acts, the more tactics the attackers develop.

REFERENCES.

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi, B., & Shah, T. (2022). Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges. Security and Communication Networks,2022(1),1862888.https://doi.org/10.1155/2022/1862888

Alghoul, A., Al Ajrami, S., Al Jarousha, G., Harb, G., & S. Abu-Naser,S.(2018). (PDF) Email Classification Using Artificial Neural Network. ResearchGate, 8–14. https://www.researchgate.net/publication/329307944_Email_Classification_Using_Artificial_Neural_Network.

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A.,Lee, G., Patterson, D. A., Rabkin, A., Stoica, I., & Zaharia, M.(2009). *Above the Clouds: A Berkeley View of Cloud Computing*.

Bassi, S., & Chaudhary, A. (2015). *Cloud Computing Data Security-Background& Benefits*. *6*, 34–40.

Bhavsar, V., Kadlak, A., & Sharma, S. (2018). Study on Phishing Attacks. *International Journal of Computer Applications, 182(33), 27-29.* https://doi.org/10.5120/ijca2018918286

Blanzieri, E., & Bryl, A. (2008). A survey of learningbased techniques of email spam filtering. *Artificial Intelligence Review* *29*(1), 63–92. https://doi.org/10.1007/s10462-009-9109-6

Ceci, L. (2024). *Topic: E-mail usage in the United States*. Statista. https://www.statista.com/topics/4295/e-mail-usage-in-the-unitedstates/

Foster, I., & Gannon, D. B. (2017). *Cloud Computing for Science and Engineering*. MIT Press.

Ibam, E. O., Boyinbode, O. K., & Afolabi, M. O. (2018). e-Commerce in Africa: The Case of Nigeria. *EAI Endorsed Transaction on Game- Based Learning,* 4(15), 153536 https://doi.org/10.4108/eai.5-1-2018.153536

Kim, S.-E., Jo, J.-T., & Choi, S.-H. (2015). SMS Spam Filterinig Using Keyword Frequency Ratio. *International Journal of Security and Its Applications*, *9*(1), 329–336. https://doi.org/10.14257/ijsia.2015.9.1.31

Kotsiantis, S. B., Zaharkis, I., & Pintelas, P. (2007).Supervised Machine Learning: A Review of Classification Techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, 3–24.

Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019). A stacking model using URL and HTML features for phishing webpage detection. *Future Generation Computer Systems*, *94*,27–39. https://doi.org/10.1016/j.future.2018.11.004

Liew, S. W., Sani, N. F. M., Abdullah, Mohd. T., Yaakob, R., & Sharum, M. Y. (2019). An effective security alert mechanism for realtime phishing tweet detection on Twitter.*Computers & Security*, *83*, 201–207. https://doi.org/10.1016/j.cose.2019.02.004

Mell, P., & Grance, T. (2011). *The NIST Definition of Cloud Computing*.

Mufti, T., Mittal, P., & Gupta, B. (2021). A Review on Amazon Web Service(AWS), Microsoft Azure & Google Cloud Platform (GCP) Services. *Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development,ICIDSSD 2020, 27-28 February 2020, Jamia Hamdard, New* Delhi*, India*. Proceedings of the 2nd InternationalConference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020, 27-28 February 2020,Jamia Hamdard, New Delhi, India, New Delhi, India.https://doi.org/10.4108/eai.27-22020.2303255

Olatunji, S. O. (2019). Improved email spam detection model based on support vector machines. *NeuralComputing and Applications*, *31*(3), 691–699. https://doi.org/10.1007/s00521-017-3100-y

Ora, A. (2020). *Spam Detection in Short Message Service Using Natural Language Processing and Machine Learning Techniques* [Masters, Dublin, National College of Ireland]. https://norma.ncirl.ie/4286/

Petersen, L. N. (2018). The ageing body in Monty Python Live(Mostly). European Journal of Cultural Studies, 21(3),382–394.

https://doi.org/10.1177/1367549417708435

RENO, R. (2025). Cloud Market Jumped to $330 billion in 2024 – GenAI is Now Driving Half of the Growth |Synergy Research Group. https://www.srgresearch.com/articles/ cloudmarketjumped-to-330billionin-2024genai-is-now-driving-half-ofthegrowth

Richter, F. (2025, February 27). Infographic: Amazon and Microsoft StayAhead in Global Cloud Market. Statista Daily Data. https://www.statista.com/chart/18819/worldwide-marketshareofleadingcloudinfrastructure-service-providers

Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017). Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets. *IOP Conference Series: Materials Science and Engineering*, *226*(1), 012091.https://doi.org/10.1088/1757-899X/226/1/012091

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019).Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345– 357. https://doi.org/10.1016/j.eswa.2018.09.029

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2021). Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey. *Procedia Computer Science*, *189*, 19–28. https://doi.org/10.1016/j.procs.2021.05.077

Yang, H., & Tate, M. (2012). A Descriptive Literature Review and Classification of Cloud Computing Research. *Communications of the Association for Information Systems*, *31*(1). https://doi.org/10.17705/1CAIS.03102

Yoo, S.-K., & Kim, B.-Y. (2019). The Effective Factors of Cloud Computing Adoption Success in Organization. *The Journal of Asian Finance, Economics and Business*, *6*(1), 217–229. https://doi.org/10.13106/jafeb.2019.vol6.no1.217