**article topic**: **RECENT DEVELOPMENTS IN SOCIAL SPAM DETECTION AND COMBATING**

**TECHNIQUES: A SURVEY**

**Reframed TOPIC: RECENT DEVELOPMENTS IN THE USE OF MACHINE AND DEEP LEARNING MODELS IN SOCIAL SPAM DETECTION: A SURVEY**

**By**

**OLAYINKA MOHAMMED KOLAWOLE**

**239074088**

**EXECUTIVE SUMMARY**

This survey explores the cutting-edge integration of machine learning (ml) and deep learning (dl) models in the detection and combat of social spam across various digital platforms. The pervasive issue of spam in digital communications, ranging from email to social networks, has necessitated the development of advanced detection techniques. With the rapid evolution of spamming methods, traditional approaches have become insufficient, prompting a shift towards employing ml and dl models for more effective spam identification and mitigation. The research delves into various domains where spam is prevalent, including email, blogs, social media comments, and reviews, showcasing the application of these computational models to enhance spam detection accuracy and efficiency.

# CHAPTER ONE

# INTRODUCTION

**1.0 Background to study**

In the era of information technology, spam also known as electronic spam describes the undesired or unsolicited messages transmitted or received through electronic means such as email, instant messaging, blogs, newsgroups, social networks, web searches, mobile phones, and more. These messages are typically sent for advertising, phishing, spreading malware, or misinformation (wikipedia, 2023). As the definition suggests spam is designed for malicious purposes and often serves as a profitable yet deceitful source of income for certain individuals or organizations, someone who engages in sending these spam messages is commonly referred to as a spammer (Chakraborty et al., 2016).

Spammers can disseminate spam through different media channels. Several categories of spam exist, based on the distribution method, including email, mobile devices, web forms, comments, SEO, social networking, messaging, and trackbacks. Email spam is the most prevalent form, with approximately 45% of daily emails being spam (Spamlaws, 2023). The emergence of Machine and Deep Learning models have played an important role in understanding and countering the adaptive strategies of spammers due to their ability to learn from large datasets and identify patterns that are not immediately obvious to a human analyst. This research will span across different social platforms where Machine and Deep Learning models have been applied, exploring their efficiency in different spam contexts ranging from emails to social media platforms.

The increase of spam on social media has become so widespread that it negatively impacts the overall user experience on the targeted social platforms, the primary cause of this increase is driven by the economic incentives and anonymity provided by social media platforms which allows spammers to operate with minimal risk and investment. Therefore, these platforms get associated with various scams and fraudulent schemes which includes phishing attacks, where spammers seek to gain financially by acquiring sensitive information such as credit card details and login credentials (Chu et al., 2012).

The spread of these spams across the social platforms has led to them being compromised, significantly deteriorating their integrity and user experience. One of the notable effects of social spam is the decrease in user engagement and trust users once had with the platform. When users get bombarded with spam messages, they become frustrated and stop using the platform, thereby reducing the activity of the user. Social spam also facilitates the spread of misinformation, which can have adverse effects on public opinion and create chaos in a society. During critical times such as elections or public health crises, misinformation propagated through spam can mislead the public and influence outcomes in an event (Allcott & Gentzkow, 2017).

**1.1 Statement of Problem**

Traditional spam detection methods, such as rule-based filters, have struggled to keep pace with the sophisticated tactics used by spammers. The exploration of Machine Learning and Deep Learning models offer promising alternatives by adapting to new spamming behaviors, techniques and more effectively identifying complex patterns. This research seeks to evaluate the current capabilities and limitations of these technologies in spam detection and addressing critical gaps across different social platforms. This gaps span across different social platforms, particularly when comparing their effectiveness for example the Machine or Deep Learning Technique that works on email spam may not necessarily be effective on Facebook spam or X spam and vice versa, so this research aims to fill this gap by providing a detailed comparative analysis of recent advancements in Machine and Deep Learning techniques and their applications across different social media platforms for spam detection.

**1.2 Aim and Objectives**

**1.2.1 Aim**

The aim of this research is to critically review and analyze the application of machine learning and deep learning models in detecting and mitigating social spam across various digital platforms.

**1.2.2 Objectives**

* To examine the current state of machine learning and further delve into the use deep learning models in spam detection, offering a comprehensive analysis of the most recent approaches for identifying spam on social media platforms.
* To assess the efficacy of these methods, a critical evaluation of performance metrics will be analyzed
* To investigate available solutions and propose suggested solutions and enhancements to existing spam detection techniques, focusing on refining strategies that integrate recent advancements in machine learning and artificial intelligence.

**1.3 Approach**

A theoretical approach to identifying and analyzing recent literature on Machine Leaerning and Deep Learning model applications in combacting social spam detection, this involves the process of combining keyword-based searches,inclusion of seminar and recent papers, and qualitative analysis of identified litetratures.

* Starting with keyword-based search on academic databases with the use of specific keywords related to Machine Learning, Deep Learning models and Social Spam Detection.
* Expansion of the literature base by reviewing set of related articles to ensure a comprehensive understanding of the field.
* Extraction of key information from each selected papers, such as objectives, methodologies, findings, and the effectiveness of different spam detection techniques within their application context.

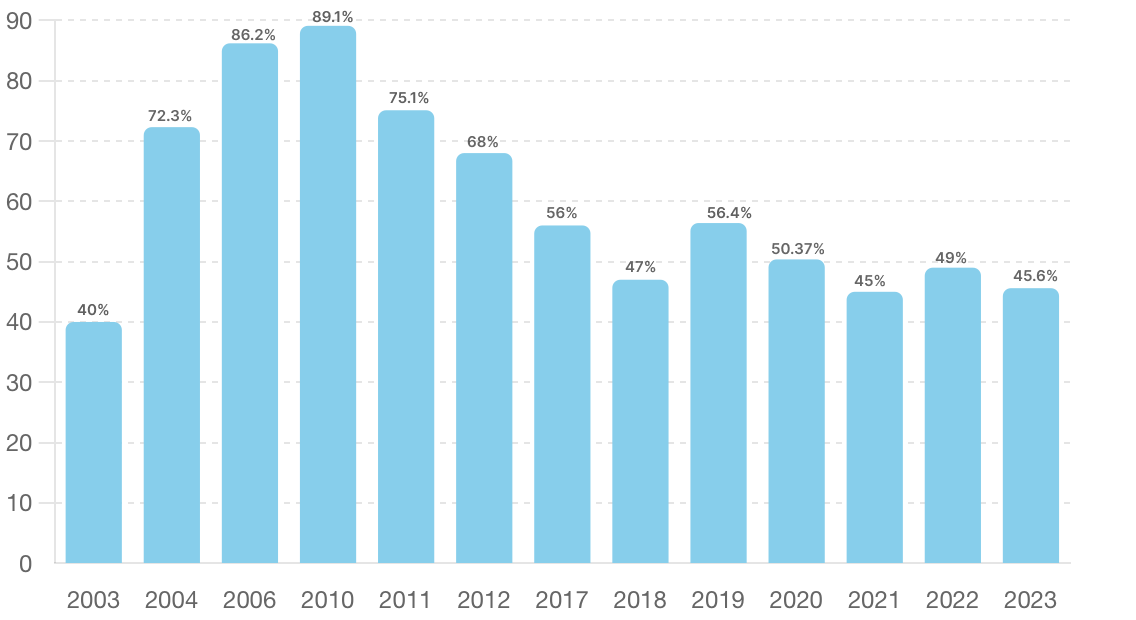
**1.4 Limitations of the Study**

Acknowledgement of any potential limitations in this study, such as the rapidly evolving nature of the field which might surpass the scope of my review, as well as the difficulties in comparing studies due to their varying methodologies and metrics across different social platform or use-case.

# CHAPTER TWO

# LITERATURE REVIEW

**2.0 History of Spam**

Spam can be traced as far back to the early 1990s, with the first recorded instance being a mass email sent to 600 recipients on ARPANET in 1978. However, it was not until the commercialization of the internet that spam became a significant issue, primarily in the form of mass unsolicited emails aimed at advertising (Templeton, 2015). With rise of social media and other digital communication platforms, spammers have diversified their method by utilizing specific techniques customized to the various social media environments they are targeting, introducing new spamming methods such as phishing scams, malware, and deceptive advertising. This shift necessitated the development of more sophisticated detection techniques to address the growing complexity of spam (Edwards & Lutz, 2014).The fight against spam has witnessed significant technological advancements, evolving from simple rule-based filters to sophisticated Machine Learning (ML) and Deep Learning (DL) algorithms along with their variants capable of identifying and mitigating spam with high precision. Despite the significant advancement witnessed in the fight against spam, it still accounted for 47.3% of global email traffic, with an estimated 122.3 billion spam emails sent daily (Statista, 2021). **Figure 2.1Global spam volume as percentage of total e-mail traffic from 2003-2023 (Statista)**

Initially spam was managed through basic rule filtering, these rules are specifically designed to identify and filter messages that meet the specific criteria which indicates spam. For example, rule-based filters work by filtering or blocking content containing specific keywords such as free, promo. According to (Blanzieri, & Bryl, 2008), the efficacy of rule-based filtering lies in its straightforward and uncomplicated method of detecting spam. These systems were simple to set up and understand, and they successfully prevented spam without requiring sophisticated computational resources. However, rule-based filters also have significant limitations. They require constant updating as spammers adapt to bypass static rules. This creates a maintenance overhead as new rules must be continually crafted, and old ones revised. So, this marked the introduction of Machine Learning Algorithms introducing the ability to learn from data and adapt to new spamming techniques rather than solely relying on predefined rules (Chakraborty, 2016).

**2.1 Literature Review**

As soon as Email was realized to be integral to various sectors such as business and education, the issue of spam email keeps rising exponentially daily, presenting a significant challenge for email and Internet of Things (IoT) service providers. In terms of detection and filtration a machine learning technique a survey was present by (Ahmed et al., 2022) on spam filtering on emails and IoT platforms. In details they used a variety of algorithms to enhance spam filtering effectiveness, including Naive Bayes, Decision Trees, Neural Networks, and Random Forest.

As discussion on spam detection in social network keeps progressing (Chaudhry et al., 2020) using Machine Learning approach talked about how proliferation of information and the accessibility of user data can potentially attract spammers, rendering social networks vulnerable to spam attacks, so a machine learning algorithm was proposed in their study for spam detection on social networks. They specifically SVM as the classifier which involves a two-phase process: the training phase and where features of dataset are calculated as weighted values and used to train the SVM, and testing phase where test data features are evaluated against the trained model to identify spam. Due to the diversity of spams in different socail media platofrms (Binsaeed et al., 2020) presented a survey on addressesing the detection of malicious activities on detecting spam in twitter microblogging, they developed a novel approach based on domain popularity to identify spam in tweets.

The rapid growth online social networks (OSNs) have made it difficult in Identifying spam. Trying to identify spams on these platforms is challenging due to the complex nature of the task and the skewed distribution of spam and legitimate content (ham), which often benefits spammers by making our devices vulnerable, (Sumathi & Raja, 2023) presented a brief survey on account of using different machine learning algorithms to combat spam in the growing online social network In their research a diverse array of machine learning algorithms including Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), XGBoost (XGB), and a Voting Classifier (VC) are utilized. This ensemble of methods is designed to address the imbalanced data distribution between spam and legitimate content, enhancing the detection accuracy significantly. Another work on social spam detection by (Rao et al., 2021) discusses how AI and deep learning enhanced Deepfake (text, image, and video) spam and their counterstrategies. They also engaged in thorough discussions about current challenges and upcoming issues like the detection systems' robustness, scalability, access to real-time datasets, spammers' evasion tactics, coordinated inauthentic behavior, and adversarial attacks on ML-based spam detectors.

Twitter currently known as X is highly regarded for enabling users to share ideas, discuss social issues, access news, and maintain connections with family and friends. However, its significant popularity has also made it a prime target for spammers, in recent developments, deep learning-based approaches have been introduced for identifying spammers on Twitter. (Manasa et al., 2021) presented a survey that examines and compares various existing methods for detecting spam on twitter. They specifically focus on the utilization of deep learning techniques, including Word Vector techniques and binary classifiers, to enhance the interpretability and accuracy of spam detection, highlighting that while some techniques have shown promising results, others have negatively impacted the effectiveness of the detection process. (Zhao et al., 2020) discussed about the growth of social media and how traditional approach has helped mitigate this risk, primarily focusing on feature-based or propagation-based detection. They further emphasized on how a semi-supervised graph embedding model leveraging a graph attention network can be used to enhance detection of spam in social networks. This model operates by constructing a detection framework that aggregates user features and the relationships between users within the network. With the recent advances in spam detection.

(Zhang et al., 2023) stated that not much progress has been made in multimodal spam detection across multiple languages stating how it remains relatively under-investigated. In their research they unveiled an innovative strategy using deep learning for detecting spam that incorporates both text and document images across different languages, known as Multilingual and Multimodal Spam Detection Model (MMTD). Further emphasizing on email spam detection which is a part of social spam and ongoing efforts to fight it, combating it has become a complex task due to the persistent nature of new spam techniques used by spammers. So, the need for enhanced detection methods needs to be introduced. In response to this (Alauthman, 2020) introduced an innovative deep learning approach that combines Gated Recurrent Unit and Recurrent Neural Network (GRU-RNN). Alauthman’s methodology enhances the detection of spam emails by leveraging the sequential dependency of data, thus offering a dynamic response to the changing patterns of spam.

Karishma et al., (2020) presented a survey stating that the primary concern with spam is its potential to downlaod malicious software, which can compromise computers, smartphones, and network system, consume network bandwidth and storage, overload email servers, and lead to device vulnerabilities through spyware, phishing, and ransomware attacks to address this they proposed a deep learning spam detection method employing recurrent neural netwroks, specifically using the Bidirectional Gated Recurrent Unit (BiGRU) model. This approach leverages the capabilities of BiGRU to process data sequentially and in both directions, thus enhancing the model's ability to capture contextual information from the data.

In study presented by (Tida & Hsu, 2022) they introduce a cutting-edge universal deep learning model for detecting spam, utilizing the pre-trained Bidirectional Encoder Representations from Transformers (BERT) by Google, specifically the base uncased variant. Their research presented a novel universal spam detection model that employs the BERT base uncased models across four different datasets, showcasing its efficiency in classifying ham or spam emails in real-time scenarios. With the continuation in development of deep learning's application in social media, a study by (Xu et al., 2023) explores a deep learning model leveraging LSTM and attention mechanisms to target spam on platforms like Twitter. This approach effectively addresses the dynamic nature of spam, allowing for real-time detection and response, thereby maintaining user trust and platform integrity.

Bhavsar et al. (2023) presented a study on tackling YouTube spam comment which is the top of video content platforms. Their approach demonstrates how targeted algorithms can effectively reduce spam interactions, thus preserving content credibility and enhancing viewer experience. To further enhance the fight against email spam, which is at the forefront of social spam, (Nagalikar, 2021) introduced the integration of LSTM with an attention mechanism to distinguish between spam and legitimate emails more effectively. This method not only improves accuracy but also adapts to the evolving nature of spam tactics, ensuring robust defense mechanisms are in place.

In a survey conducted by (Hussain et al., 2020). They explored the identification of spam reviews in online platforms through behavioral analysis of spammers. The research highlights the exploitation of six spammer behavioral features, demonstrating an enhanced detection method which uses weight calculations to classify reviews with 84.5% accuracy. Their work underlines the significance of recognizing spam activity to maintain the integrity of user-generated content on the internet​​.

Loucif’s (2024) proposed a hybrid deep learning model that combines Principal Component Analysis (PCA) and Convolutional Neural Networks (CNN) for effective spam detection on Twitter. This model addresses the challenge of traditional machine learning models struggling with the platform’s noise and brevity, achieving higher precision (94.91%), recall (96.76%), and F-score (95.83%) compared to other baseline models. The innovative approach leverages the strength of PCA in reducing dimensionality and noise, enhancing the CNN’s performance in spam detection​​. In a study by (Sadiq et al., 2016) they presented a survey thatfocuses on the detection of deepfake content on social media platforms, utilizing advanced deep learning methods.

As social network platforms like Twitter and Sina Weibo grow in popularity, so is the rate at which spammer try to spread misinformation and illegal content. To overcome these challenges, (Shen et al., 2022) presented a survey on spam detection using CNMFSD, which innovatively leverages both the content shared by users and their interaction patterns. This method was rigorously tested on a real-world dataset from Twitter, with the findings indicating a significant enhancement in detection efficacy over existing baseline models.

In a survey by (Reddy et al., 2023) they employed a variety of machine learning algorithms to refine email spam detection. Their work underscores the effectiveness of integrating multiple algorithms to improve the detection process, catering to the nuances of email communication. The influence of online public reviews on consumer behavior, especially in product purchases or service selection, cannot be overstated. With the rise of fake reviews aimed at either promoting or undermining products or the reputation of organizations, detecting spam reviews has become a critical area of research. (Neisari et al., 2021) proposes an innovative model for identifying fake reviews, leveraging linguistic features. Their approach combines unsupervised learning through self-organizing maps (SOM) with the analytical power of convolutional neural networks (CNN) for review classification.

Also, on review spam a type of social spam (Shahariar et al., 2022) developed methods in identifying them by incorporates a blend of deep learning techniques Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) alongside classic machine learning classifiers like Naïve Bayes (NB), K Nearest Neighbor (KNN), and Support Vector Machine (SVM). By leveraging both labeled and unlabeled data, their approach effectively enhances the detection of deceptive reviews, offering a robust solution to a pressing online issue. In 2013, Tan et al created a robust spam detection model. Initially, they introduced an automated spam detection framework called SD2, which uses Sybil defense mechanisms and significantly improves on current methods by taking social network relationships into account. To address the rise in spam attacks, they further developed an unsupervised detection system named UNIK. Unlike traditional methods that focus on identifying spammers directly, UNIK works by deliberately removing non-spammers from the network, effectively narrowing down the suspects and improving detection accuracy.

An advanced social spam detection method was Introduced by (Xu et al., 2021), their approach uses the Bi-LSTM neural network framework. This method employs ALBERT to convert texts from social media into word vectors, which are subsequently processed by the Bi-LSTM layer for feature extraction. The superiority of their model was confirmed through various performance metrics, including accuracy, precision, and F1-score, demonstrating its enhanced effectiveness in spam detection compared to previous models.

As researchers continued to use deep learning models to tackle spam (Zavrak & Yilmaz, 2022), introduces an innovative method for detecting email spam, leveraging a synergistic approach that combines convolutional neural networks (CNNs), gated recurrent units (GRUs), and attention mechanisms. The technique focuses on specific segments of email content during training, utilizing convolution layers for deep feature extraction through a hierarchical structure.

Olatunji, in 2019 created a spam filtering tool using support vector machine (SVM) and extreme learning machine (ELM) algorithms, employing a standard dataset for developing the spam detection model. The SVM achieved an accuracy of 94.06%, while the ELM reached an accuracy of 93.04%, showing a marginal 1.1% improvement in SVM's performance over ELM. This suggests that, given the slight accuracy difference, the ELM spam detector is preferable to the SVM in scenarios where detection speed is crucial, such as in real-time systems. An alternative approach to spam detection was introduced by. (Akhmiri & Haroonabadi, 2016) by employing a fuzzy decision tree and the Naïve Bayes algorithm. They utilized the baking voting algorithm to identify patterns in spam behavior, acknowledging that clear spam characteristics are often absent in real-world scenarios. Instead, they relied on rational and neutral cross-linking degrees to describe or explain these characteristics. The decision trees were built using fuzzy Mamdani rules to classify emails as either spam or ham.

As the investigation of email spam keeps progressing (Kumar et al., 2020) detection through various machine learning algorithms. Their study examines these machine learning methods and their application to datasets. Among the algorithms tested, the one delivering the highest precision and accuracy for email spam detection was identified. The Multinomial Naïve Bayes algorithm was found to yield the best results, although it has limitations due to class-conditional independence, which occasionally leads to misclassification. Ensemble models ranked next, producing reliable results. The proposed system can only detect spam based on the email body.

In 2016, Xu et al. devised a method to detect spam in online social networks, focusing on the cross-platform nature of spam between different networks. Using Twitter, they collected 1,937 spam and 10,943 ham tweets, as well as 1,338 spam and 9,285 ham posts from another platform. In Twitter spam detection (TSD), 75.6% of spam tweets contained URLs, while 24.4% contained specific words. Of the 10,942 ham tweets, 62.9% had URL links and words, while 37.1% contained only words. In the Facebook spam detection (FSD) context, 32.8% of spam posts had web links, while the remaining 67.2% contained only words. Of the 9,285 ham posts, 95.1% included web links, and 4.9% only words. Xu et al. used the top 20 feature words from Facebook and Twitter spam datasets, splitting each into training and testing datasets for machine learning classifiers like Naïve Bayes, random forest, logistic regression, and others. They then trained classifiers using combined datasets from both platforms. This combined approach showed higher accuracy than any single dataset alone

Guo et al., 2020 introduced a collaborative neural network-based spam detection method for IoT applications, called Cospam. They considered feature sequences from user behavior and speech across various timestamps. Cospam's collaborative model included Bi-AE, GCN, and LSTM models to identify user behavior. Experiments revealed that Cospam achieved 5% better accuracy than existing approaches but required more time due to the complexity of parameters.

Furthermore, on the use of deep learning techniques (Makkar and Kumar, 2020) proposed a deep learning model to detect web spam in IoT environments. This system improves search engine detection of web spam using rank scores calculated by the search engine. It leverages deep learning features and uses an LSTM model that was previously applied to problems like weather forecasting. Their framework was tested against ten different machine learning models using the WEBSPAM-UK 2007 dataset. The preprocessing used a unique technique called Split by Oversampling and Train by Underfitting. The proposed model achieved an accuracy of 95.25%, and after optimization, it reached 96.96%.

# CHPATER THREE

As discussed in Chapter One, the proliferation of social spam is driven by various factors, including economic incentives, low entry barriers, and the inherent vulnerabilities of digital platforms. This chapter delves into specific types of social spam and the corresponding spamming techniques, linking them to the Machine and Deep Learning techniques designed to detect and mitigate their impact.

**3.0 Types of Social Spam and Corresponding ML/DL Detection Techniques**

**3.1 Phishing Spam**

Phishing spam is a type of social engineering attack that uses a false identity in electronic interactions to gain sensitive data, including credit card numbers, usernames, and passwords. These attacks, which are usually conducted through email, social media, SMS, or instant messaging, entail delivering misleading communications that seem to be from reliable sources **to** fool victims into divulging sensitive personal information or downloading malicious software (Hadnagy & Wilson, 2010).

**3.1.1 Techniques Used in Phishing Attacks**

* Spear Phishing: Uses information acquired to develop persuasive content that appears highly relevant to the target, targeting certain people or organizations with customized messages (Jagatic et al., 2007).
* Hidden URLs: This type of URL masking technique uses words such as Click Here to direct users to a malicious website (Garera et al., 2007).
* Link Manipulation: This technique, which sometimes involves tiny misspellings or the usage of a subdomain, uses visually identical URLs to trick people into believing the phishing site (Garera et al., 2007).

**3.1.2 Economic Incentives as a Driving Force Behind Phishing Spam**

Phishing attack which is category of social spam, pose a serious risk, and the motivation of attackers is largely driven by financial gain. The purpose of these malicious operations is to obtain credit card numbers, usernames, passwords, and other sensitive financial and personal information illegally so that it can be utilized for financial benefit. In this article, I'll examine the financial incentives that drive phishing attacks and explain how these incentives contribute to the growth of these kinds of online spam.

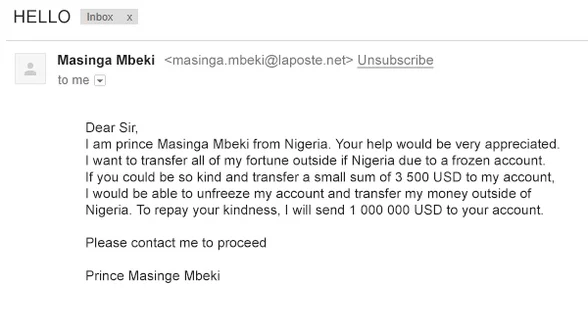
**3.1.3 The Nature of Economic Incentives in phishing**

* Direct Monetary Gain: Direct monetary gain is the main source of motivation for those who carry out phishing attacks. Cybercriminals can carry out illicit activities or sell the information on the dark web by gaining access to bank accounts, credit card details, or other financial data. (Anderson et al., 2019) stated that direct financial theft is still the easiest and fastest way for scammers to profit from phishing attempts, which is why it is a popular option.
* Data Harvesting: Phishing attacks often involve harvesting large volumes of personal data, which can be sold on underground markets. The data collected can include email addresses, passwords, social security numbers, and more. (McCoy et al., 2017) discussed on how the trade in stolen data represents a significant part of the cybercrime economy.

**3.1.4 Available Machine Learning Techniques to Combat Phishing**

Random Forests for Phishing Detection

A machine learning approach utilizing an ensemble of decision trees to enhance the detection of phishing emails. This method efficiently processes a feature set extracted from email data, which includes characteristics such as IP-based URLs, HTML content, and mismatched link text, among others. By generating multiple decision trees and aggregating their outcomes, Random Forests achieve high accuracy and reliability, effectively minimizing both false positives and false negatives. This robust classification system is particularly adept at adapting to the evolving tactics of phishing, ensuring reliable security measures against various forms of email-based fraud(Akinyelu & Adewunmi, 2014).



**Figure 3.1 An example of Phishing email sent to a User**

**3.2 Spam Bots**

Spam bots are automated programs created to imitate human behavior on the internet. They produce unwanted and repetitive messages to advertise goods, disseminate false information, or influence attitudes and trends on online platforms. To disseminate their messages more efficiently and reach a wider audience, these bots might pose as actual users on social media networks, forums, and email systems (Ferrara et al., 2016).

**3.2.1 Techniques Used by Spam Bots**

* Social Media Bots: These bots create fake profiles or hijack legitimate accounts to spread misinformation, amplify specific hashtags, and manipulate public opinion. They often mimic human activity by liking, sharing, or commenting on posts (Chu et al., 2010).
* Email Spam Bots: Automated programs send unsolicited bulk emails (spam) to massive lists of recipients. They frequently promote counterfeit products, phishing links, or malware attachments (Stringhini et al., 2010).
* Web scraping bots: These programs gather email addresses and other contact details from social media and websites that are accessible to the public in order to create mailing lists for upcoming spam operations (Bhat & Abulaish, 2013).
* Comment Spam Bots: Automated remarks are placed on blogs, news sites, and forums to advertise services, goods, or propaganda. These comments frequently contain links to other websites (Jindal & Liu, 2008).

**3.2.2 Social Platform Vulnerability as Motivation for Spam Bots**

Platform vulnerabilities refer to the weaknesses in social media networks, blogs, email systems, and other online services that can be exploited by spam bots to conduct malicious activities like spreading unsolicited messages or manipulating opinions. These vulnerabilities create an ecosystem conducive to exploitation by spam bots due to their automated nature and ability to mimic human behavior. This overview explores how platform vulnerabilities serve as a significant motivation for spam bots.

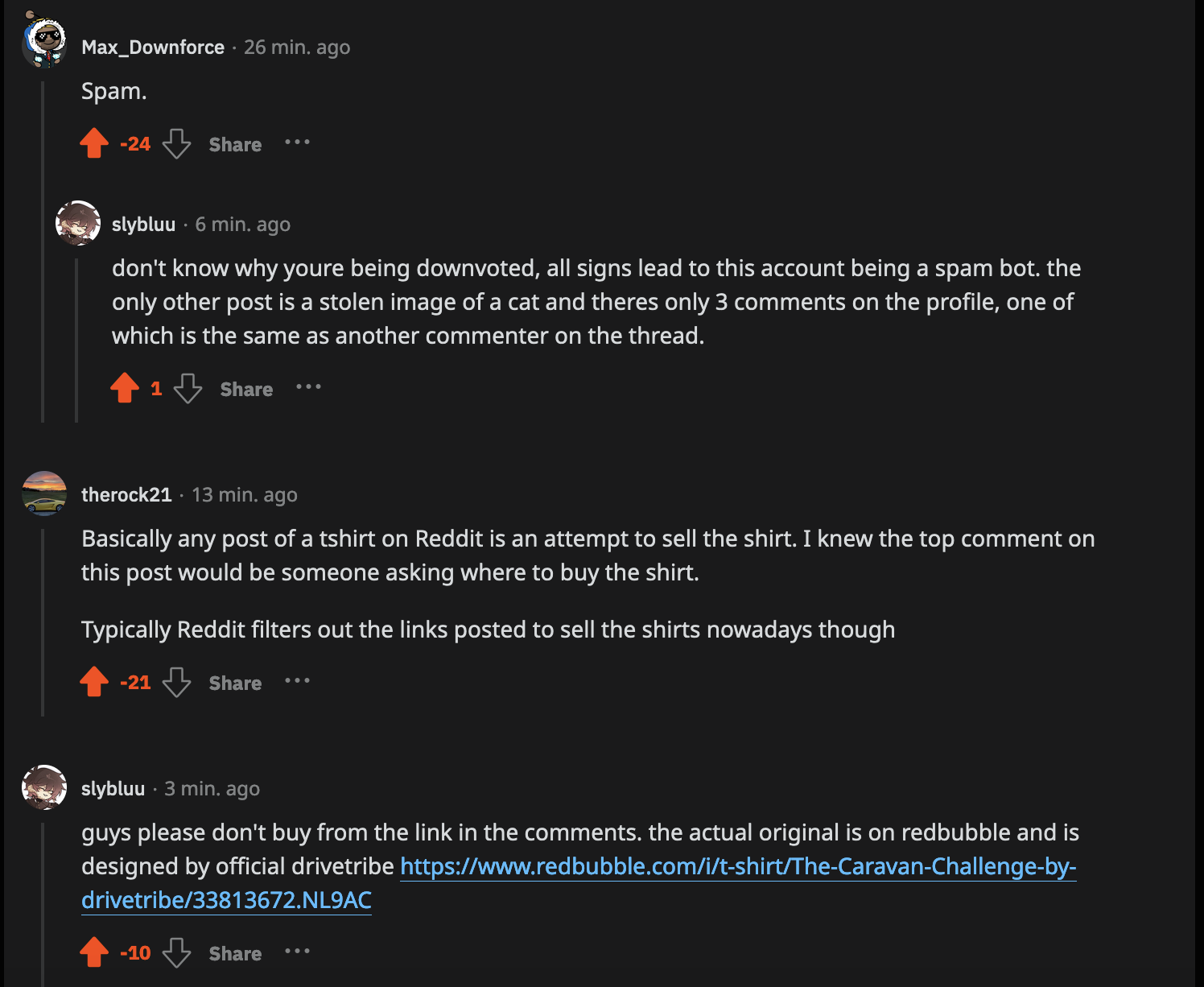
**Types Of Social Platform Vulnerabilities and How They Exploited**

* Exploiting weak authentication system: Many platforms have insufficient authentication mechanisms, making it easier for spam bots to create and operate multiple fake accounts without detection. Weak or absent CAPTCHA systems, simple password requirements, and lack of multi-factor authentication are common vulnerabilities that spam bots exploit. (Cresci et al., 2017) show that social spambots can bypass basic user verification, leveraging this to generate realistic-looking profiles that are often difficult to distinguish from genuine users.
* API Manipulation: Spam bots often take advantage of poorly secured application programming interfaces (APIs) that allow them to automate actions such as posting messages or creating accounts. APIs are essential for the scalability and interoperability of modern platforms but can also serve as a significant security gap if not properly secured. (Canali et al., 2011) found that some social media APIs have flaws that allow bots to access more data than intended, enabling mass data collection for phishing or spamming purposes.

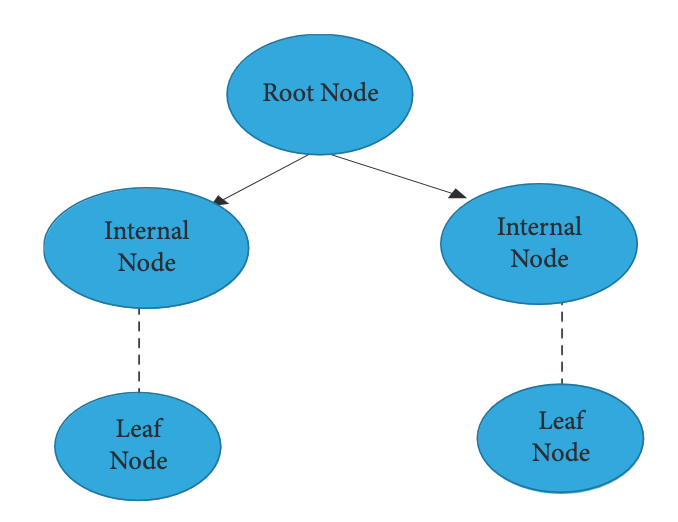
**3.2.3 Available Machine Learning Techniques to Combat Spam Bots in Social Platforms**

Random Forests

An ensemble method that analyzes over a thousand features, such as user metadata, network patterns, content characteristics, and activity time series, to distinguish legitimate users from social bots. By leveraging the Gini coefficient for optimal feature selection, Random Forests aggregate predictions from multiple decision trees to accurately detect a range of bot behaviors, from simple spammers to sophisticated self-promoters. This adaptable approach, validated through comprehensive training and benchmarking, offers robust spam bot detection by continuously refining and improving based on emerging data patterns (Varol et al., 2017).



**Figure 3.2 An example of spam bot in comment section**

****

**Figure 3.3 Structure of Random Forest Decision Tree (Ahmed, 2022)**

# CHAPTER FOUR

**4.0 Cross-Platform Spam Techniques**

Detecting spam across different social platforms remains a significant challenge due to the inherent diversity of formats, user behaviors, and platform-specific features. For instance, social media platforms like Twitter and Facebook deal with text posts, comments, and shared links, while email systems handle direct messages and attachments. Furthermore, different platforms attract distinct types of spam, ranging from phishing and scam emails to bot-generated comments and fake reviews (Ferrara et al., 2016). This variability complicates detection efforts, necessitating advanced methods that can adapt to diverse types of spam across various platforms.

**4.1 Deep Learning Techniques for Cross Platform Compatibility**

**4.1.1 Convolutional Neural Networks (CNNs) For Detecting Spam and Cross-Platform Adaptability**

CNNs is a deep learning model initially designed for image recognition, but they have been highly effective for analyzing text, particularly in spam detection. By using convolutional layers, CNNs can identify patterns like n-grams and word sequences that may signal spam or malicious content. This capability allows them to detect deceptive language or unusual behaviors in messages that could be indicative of phishing, fraudulent links, or other harmful activity. This adaptability makes CNNs well-suited for cross-platform spam detection, handling diverse data from emails, social media, and messaging apps. By processing large volumes of text efficiently, they can discern subtle features distinguishing legitimate content from spam, reducing false positives and improving overall detection.

**CNN Model**

* Input Layer: Word vectors like GloVe or word2vec are frequently used to tokenize and convert text data (such as email bodies or social media postings) into embeddings.
* Convolutional Layers: Filters scan the input text in order to identify significant n-gram features, such as certain terms or phrases that may be signs of spam.
* Pooling Layers: By reducing dimensionality, pooling layers enable the model to concentrate on its most important properties.
* Fully Connected Layer: Compiles characteristics and generates a probability score that indicates if the text input is legitimate or spam.
* Softmax Layer: Provides class probabilities based on the output score.

**Applications Across Social Platforms**

Reddy & Ahila, (2022) presented a paper on Email spam detection using Convolutional Neural Networks (CNN) algorithms. They compared the performance of these classifiers, utilizing a Kaggle dataset containing both spam and non-spam emails. In their approach, CNN was implemented with convolutional layers for feature extraction and classification, The study used a sample size of 20, with 70% of the data for training and 30% for testing. They assessed accuracy, standard deviation, and statistical significance between the two algorithms. The CNN achieved a mean accuracy of 91.18% The study demonstrates that the proposed CNN-based approach significantly detects spam emails, with better precision and robustness.

Abdelwahab &Mostafa, (2022) presented a paper on detecting Twitter spam using a combination of convolutional neural networks (CNNs) and other deep learning models. They devised a three-layer framework that combines tweet-level detection with statistical feature detection to build a robust spam classification system. The first layer is a fast filter mode classifier, which initially identifies tweets that are likely to be spam. These tweets are then paraphrased to generate new versions with the same meaning but different wording The third layer applies CNNs, to analyze both content and statistical features. The CNN-based approach proved highly effective in identifying patterns and features that distinguish spam tweets from legitimate ones, contributing to the robust performance of the overall spam detection system​​.

**Table 4.1: Comparison of Effectiveness Across Platforms**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Platform** | **Type of Spam** | **Detection Accuracy** | **Reference** |
| CNN | Email | Phshing Emails | 91.8% | Reddy & Ahila |
| CNN | Twitter | Tweet Spam | 95% | Abdelwahab &Mostafa |

# 

**Figure 4.1 A CNN Model Structure (Xiang, 2020)**

# 4.1.2 Long Short-Term Memory (LSTM) For Detecting Spam and Cross Platform Adaptability

A Long Short-Term Memory (LSTM) network is a type of recurrent neural network (RNN) designed to handle sequential data and retain information over long periods, making it suitable for detecting spam in textual content such as emails and social media posts. It is structured with specialized memory cells that help address the vanishing gradient problem, a common challenge in traditional RNNs. These memory cells contain input, output, and forget gates, which manage the flow of information, allowing the model to maintain relevant context across lengthy text sequences. In the context of spam detection, LSTM networks excel at identifying patterns in email headers, bodies, and metadata that can differentiate between spam and legitimate messages. They can learn and adapt to evolving spamming techniques, enabling the identification of subtle or context-based features indicative of malicious content

**LSTM Model**

* Memory Cell: The central unit that stores information over time, allowing the network to maintain a long-term state.
* Input Gate: Controls the amount of new information from the current input that will be added to the memory cell.
* Forget Gate: Determines how much of the existing information in the memory cell should be retained or forgotten.
* Output Gate: Controls how much of the memory cell state should be passed to the next layer as the output.
* Cell State Update: Combines the forget and input gates to update the cell state. New state = (Old state \* Forget gate) + (New input \* Input gate).
* Hidden State: The hidden state is the output of the current LSTM unit, passed to the next time step as input, and used in the final decision-making process.

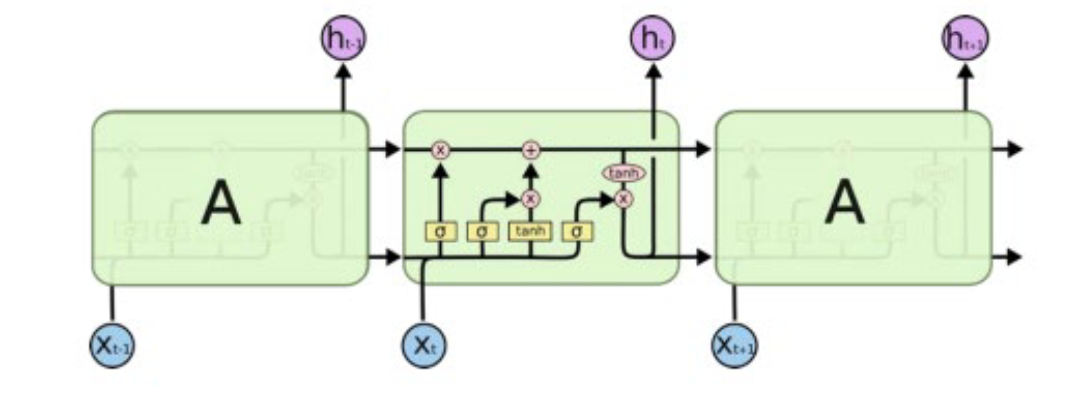
**Applications Across Social Platforms**

Xiang et al., (2020) present a paper on spam detection in product reviews using Long Short-Term Memory (LSTM) networks with multi-entity temporal features. They compared the performance of their Multi-Entity Temporal Feature-Based Spam Detection (MTFSD) model against other established spam detection methods using the Yelp dataset containing both spam and non-spam reviews. In their approach, LSTM was employed to analyze temporal patterns within reviews across user and product entities, while Convolutional Neural Networks (CNNs) were utilized to extract text features. The study used a sample size of 6,883,290 reviews, with 70% of the data for training and 30% for testing. They evaluated the accuracy, precision, recall, and F1-score of the MTFSD model compared to baseline models. The MTFSD model achieved a mean accuracy of 91.8%.

Busyra & Girsang, (2024) presented a paper on applying Long Short-Term Memory (LSTM) networks to detect spam submitted through web forms on government ministry websites. They used multilingual data in both English and Indonesian to develop their model. In their approach, the LSTM network processes textual data that has undergone preprocessing steps such as tokenization, stemming, and stop word removal. This is followed by word embedding using Word2Vec to convert words into numerical vectors. The study concluded that the proposed LSTM model effectively detects spam messages in web forms, particularly on government ministry websites, and can handle both English and Indonesian languages. This paper also emphasizes the importance of data augmentation and preprocessing techniques in enhancing the accuracy and robustness of spam detection.

**Table 4.2: Comparison of Effectiveness Across Platforms**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Platform** | **Type of Spam** | **Detection Accuracy** | **Reference** |
| LSTM | Yelp | Review Spam | 91.8% | Xiang et al., (2020) |
| LSTM | Web | Web Form Spam | 95% | Busyra & Girsang, (2024) |



**Figure 4.2 An LSTM model structure (Busyra et al, 2024)**

# 4.1.3 Transformer Model for Detecting Spam and Cross Platform Adaptability

Transformer models represent a groundbreaking deep learning architecture designed for sequential data tasks. Unlike Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, transformers process data in parallel using a self-attention mechanism instead of processing sequences step-by-step. This parallelization allows transformers to achieve significantly higher efficiency and performance in tasks like machine translation, text classification, and spam detection (Vaswani et al., 2017)

**Transformer Architecture**

* Input Layer: Accepts sequential data (e.g., text sentences) and applies positional encoding to represent word order.
* Encoder: Consists of multiple identical layers with two sub-layers each: multi-head attention and feed-forward networks.
* Decoder: Similar to the encoder but with an additional sub-layer that attends to the encoder's output.
* Output Layer: Produces predictions based on the decoder's final state (e.g., translated text, classification of spam, or legitimate).

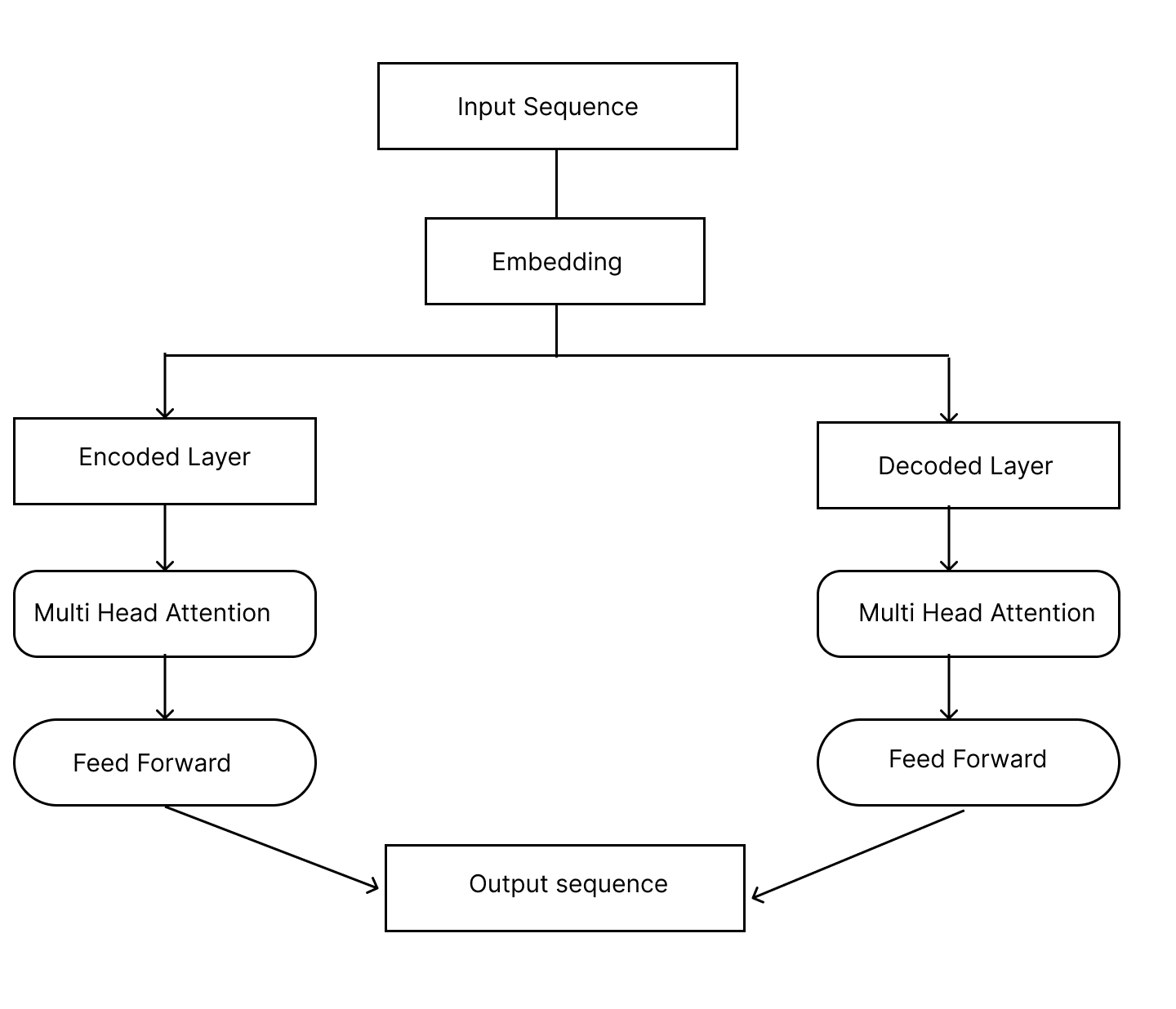
**Applications Across Social Platforms**

Jamal et al., (2023) present a paper on detecting phishing, spam, and ham emails using an Improved Phishing Spam Detection Model (IPSDM) based on transformer models. They compared the performance of their IPSDM model, which fine-tunes DistilBERT and RoBERTa, against baseline transformer models on phishing and spam detection datasets. In their approach, both DistilBERT and RoBERTa were fine-tuned using a dataset that combined phishing, spam, and ham emails from multiple sources. Adaptive Synthetic Sampling (ADASYN) was used to balance the datasets and address class imbalances. They implemented learning rate scheduling, batch size adjustments, and early stopping to optimize the training process. The study demonstrates that the transformer based IPSDM approach effectively identifies phishing and spam emails, offering superior detection accuracy and robustness.

Ilias et al., (2023) present a paper on detecting social spambots on Twitter using a multimodal transformer-based model. They compared the performance of their model, which combines TwHIN-BERT and VGG16, against existing baseline methods for identifying spambots. In their approach, TwHIN-BERT was fine-tuned to analyze textual data from Twitter profiles, while VGG16 extracted visual features from digital DNA sequences created based on users' behavioral patterns. A cross modal attention mechanism captured the relationship between text and visual features, and a Gated Multimodal Unit (GMU) adjusted the importance of each modality to the final classification. The approach achieved an accuracy of up to 99.98% on the Cresci '17 dataset, significantly outperforming previous detection models

**Table 4.3: Comparison of Effectiveness Across Platforms**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Platform** | **Type of Spam** | **Detection Accuracy** | **Reference** |
| Transfomers | Email | Email Spam | 98% | Jamal et al., (2023) |
| Transfomers | Twitter | Twitter | 99.8% | Ilias et al., (2023) |



**Figure 4.2 A Basic Transformer model structure (Zhang, 2023)**

**4.2 FUTURE TRENDS**

The future of social spam detection or any type of spam detection in general lies in the combination of various machine learning and deep learning models to form a hybrid model. A Hybrid model can leverage the strengths of both Machine Learning interpretability and efficiency with Deep Learning powerful pattern recognition this gives it the ability to enhance spam detection accuracy. By integrating advanced feature engineering, these hybrid models should be able to perform real-time spam detection, adapting to new spam tactics through continuous learning and model updating. Hybrid models would also be designed to have cross-platform compatibility ensuring effective spam detection across various social media and email platforms, thereby offering a comprehensive solution to the evolving challenges faced in social spam detection.

# CHAPTER FIVE

**CONCLUSION**

This survey of recent developments in machine learning (ML) and deep learning (DL) models for social spam detection underscores the profound impact of these technologies in combating the evolving threat of unsolicited digital content. Spammers continually adapt their techniques to exploit the vulnerabilities of digital platforms like social media and email, leading to significant challenges such as phishing attacks, spam bots, and misinformation campaigns. Traditional rule-based spam filters are becoming increasingly ineffective as spammers manipulate text, links, and behaviors to evade detection. Consequently, models like Support Vector Machines (SVM) and Random Forests offer notable improvements in detecting phishing attacks, spam bots, and other forms of malicious content across digital platforms. Their ability to learn from data has made them valuable tools in countering emerging spam tactics.

# REFRENCES

Chakraborty, M., Pal, S., Pramanik, R., & Chowdary, C. R. (2016). Recent developments in social spam detection and combating techniques: A survey. Information Processing and Management, 52(6), 1053–1073

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing, 9(6), 811-824.

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, 31(2), 211-236

Blanzieri, E., & Bryl, A. (2008). A Survey of Learning-Based Techniques of Email Spam Filtering. Technical Report DIT-06-056. University of Trento, Italy.

Templeton, B. (2015). Origin of the term ”spam” to mean net abuse.

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, Article ID 1862888

(Blanchieri & Bryl, 2008),

Chaudhry, S., Dhawan, S., & Tanwar, R. (2020). Spam Detection in Social Network Using Machine Learning Approach. In U. Batra, et al. (Eds.), REDSET 2019: Recent Developments in Social Networks. Communications in Computer and Information Science, vol. 1230, pp. 236–245. Springer, Singapore.

Binsaeed, K. A., Stringhini, G., & Youssef, A. E. (2020). Detecting Spam in Twitter Microblogging Services: A Novel Machine Learning Approach based on Domain Popularity. International Journal of Advanced Computer Science and Applications, 11(11).

Sumathi, M., & Raja, S. P. (2023). Machine learning algorithms-based spam detection in social networks. Social Network Analysis and Mining.

Rao et al., (2021)

Manasa, P., Malik, A., Batra, I., & Luhach, A. K. (2021). A Comparative Study on Twitter Spam Detection Using Deep Learning Techniques. Algorithms, Computing and Mathematics Conference.

Zhao, C., Xin, Y., Li, X., Zhu, H., Yang, Y., & Chen, Y. (2020). An Attention-Based Graph Neural Network for Spam Bot Detection in Social Networks. Applied Sciences, 10(8160).

Zhang, Z., Deng, Z., Zhang, W., & Bu, L. (2023). MMTD: A Multilingual and Multimodal Spam Detection Model Combining Text and Document Images. Applied Sciences, 13(11783), 1-21.

Karishma, S., Akila, V., & Govindasamy, V. (2020). Spam Detection using Recurrent Neural Networks. International Journal for Research in Engineering Application & Management (IJREAM), 6(1).

Tida, V. S., & Hsu, S. (2022). Universal Spam Detection using Transfer Learning of BERT Model. In Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS-2022).

Bhavsar, D., Dcruz, S., Chandekar, A., Chaudhari, M., & Patil, C. (2023). A Research on YouTube Spam Comments Detection and Deletion. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11(V),

Nagalikar, G. A. (2021). Improved Email Spam Detection Using An Integrated Approach Of LSTM And Attention Mechanism. Webology, 18(6).

Hussain, N., Mirza, H. T., Hussain, I., Iqbal, F., & Memon, I. (2020). Spam Review Detection using the Linguistic and Spammer Behavioral Methods. IEEE Access

Loucif, H. (2024). A Hybrid Deep Learning Approach for Spam Detection in Twitter. Ingénierie des Systèmes d’Information, 29(1), 117-123.

Sadiq, S., Aljrees, T., & Ullah, S. (2023). Deepfake Detection on social media: Leveraging Deep Learning and FastText Embeddings for Identifying Machine-Generated Tweets. IEEE Access.

Shen, H., Wang, B., Liu, X., & Zhang, X. (2022). Social Spammer Detection via Convex Nonnegative Matrix Factorization. IEEE Access, 10.

Reddy, A., Reddy, K. H., Abhishek, A., Manish, M., Dattu, G. V. S., & Ansari, N. M. (2023). Email Spam Detection Using Machine Learning. Journal of Survey in Fisheries Sciences, 10(1), 2658-2664.

Shahariar, G. M., Biswas, S., Omar, F., Shah, F. M., & Hassan, S. B. (2022). Spam Review Detection Using Deep Learning.

E. Tan, L. Guo, S. Chen, X. Zhang, and Y. Zhao, “Unik:unsupervised social network spam detection,” in Proceedings of the 22nd ACM international conference on Information & Knowledge Management, ACM, San Francisco, CA, USA, October 2013

Xu, G., Zhou, D., & Liu, J. (2021). Social Network Spam Detection Based on ALBERT and Combination of Bi-LSTM with Self-Attention. Security and Communication Networks, 2021, Article ID 5567991.

Zavrak, S., & Yilmaz, S. (Year). Email Spam Detection Using Hierarchical Attention Hybrid Deep Learning Method.

S. O. Olatunji, “Improved email spam detection model basedon support vector machines,” Neural Computing & Applications, vol. 31, no. 3, pp. 691–699, 2019.

H. Takhmiri and A. Haroonabadi, “Identifying valid email spam emails using decision tree,” International Journal of Computer Applications Technology and Research, vol. 5, 2016

N. Kumar and S. Sonowal, “Email spam detection using machine learning algorithms,” in Proceedings of the 2020 Second International Conference on Inventive Research inComputing Applications (ICIRCA), pp. 108–113, Coimbatore, India, 2020

Xu, W. Sun, and A. Javaid, “Efficient spam detection across online social networks,” in Proceedings of the 2016 IEEE In- ternational Conference on Big Data Analysis (ICBDA), IEEE, Hangzhou, China, March 2016.

Guo, Y. Shen, A. K. Bashir et al., “Robust spammer de- tection using collaborative neural network in Internet of thing applications,” IEEE Internet of Things Journal, vol. 8, 2020

Makkar and N. Kumar, “An efficient deep learning-based scheme for web spam detection in IoT environment,” Future Generation Computer Systems, vol. 108, pp. 467–487, 2020

Hadnagy, C., & Wilson, P. (2010). Social Engineering: The Art of Human Hacking. Wiley Publishing.

Jagatic, T. N., et al. (2007). "Social phishing." Communications of the ACM, 50(10), 94-100.

Garera, S., et al. (2007). "A framework for detection and measurement of phishing attacks." Proceedings of the 2007 ACM workshop on Recurring malcode, 1-8.

McCoy, D., Dharmdasani, H., Krebs, B., Krebs, K., & Grant, A. (2017). "Priceless: The Role of Payments in Abuse-advertised Goods." Journal of Cybersecurity, 3(1), 45-56.

Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J., Levi, M., ... & Savage, S. (2019). "Measuring the Changing Cost of Cybercrime." In Eleventh Workshop on the Economics of Information Security (WEIS).

Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of Phishing Email Using Random Forest Machine Learning Technique. Journal of Applied Mathematics, 2014, Article ID 425731.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). "The rise of social bots." Communications of the ACM, 59(7), 96-104

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" IEEE Transactions on Dependable and Secure Computing, 9(6), 811-824.

Stringhini, G., Kruegel, C., & Vigna, G. (2010). "Detecting spammers on social networks." Proceedings of the 26th Annual Computer Security Applications Conference, 1-9.

Jindal, N., & Liu, B. (2008). "Opinion spam and analysis." Proceedings of the 2008 International Conference on Web Search and Data Mining, 219-230.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race." Proceedings of the 26th International Conference on World Wide Web Companion, 963-972.

Canali, D., Balzarotti, D., & Francillon, A. (2011). "The role of web hosting providers in detecting compromised websites." Proceedings of the 20th USENIX Security Symposium, 12-15.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). "Online human-bot interactions: Detection, estimation, and characterization." Proceedings of the International AAAI Conference on Web and Social Media, 280-289.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). "The rise of social bots." Communications of the ACM, 59(7), 96-104

Abdelwahab, A., & Mostafa, M. (2022). A Deep Neural Network Technique for Detecting Real-Time Drifted Twitter Spam. Applied Sciences, 12(6407), 1-20

Reddy, Y. T. K., & Ahila, S. S. (2022). Spam Detection on Emails Using Convolutional Neural Network Classifier with K Nearest Neighbor Classifier. Baltic Journal of Law & Politics, 15(4), 104-110.

Xiang, L., Guo, G., Li, Q., Zhu, C., Chen, J., & Ma, H. (2020). Spam detection in reviews using LSTM-based multi-entity temporal features. Intelligent Automation & Soft Computing, 26(6), 1375-1390.

Busyra, R. F., & Girsang, A. S. (2024). Applying Long Short-Term Memory Algorithm for Spam Detection on Ministry Websites. Journal of System and Management Sciences, 14(2), 1-20.

Vaswani, A., et al. (2017). "Attention is all you need." Advances in Neural Information Processing Systems, 30.

Jamal, S., Wimmer, H., & Sarker, I. H. (2023). An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach.

Ilias, L., Kazelidis, I. M., & Askounis, D. (2024). Multimodal Detection of Social Spambots in Twitter using Transformers.