GAS Journal of Engineering and Technology (GASJET)

OPEN ACCESS

Volume 2, Issue 9, 2025 Journal Homepage: https://gaspublishers.com/gasjet-home/ ISSN: 3048-5800
Email: editor@gaspublishers.com

Analysis of Current Happenings in Spam Detection and Solution Approaches in Cyberspace

Idowu Omolara Anike & Dr. N. A. Nurayn

Federal University of Agriculture, Abeokuta Ogun State

Received: 29.08.2025 | **Accepted:** 21.09.2025 | **Published:** 26.09.2025

***Corresponding Author:** Fasiku, Gbenga Cornelius Ph.D.

DOI: *10.5281/zenodo.170286708*

Abstract **Original Research Article**

In recent years, spam has infiltrated electronic communication, exploiting social media platforms like Facebook, Twitter, and YouTube due to their user growth and content openness. This has led to distinct challenges in combating spam on these platforms, as their characteristics differ from email and web search engines. The ever-evolving nature of social media exacerbates this issue, demanding specialized countermeasures. Despite being a nascent field, recent endeavours to counter social spam are abundant and ongoing. This paper delves into theoretical models and practical applications, summarizing advances in social spam identification and reduction. By comparing strategies, it outlines recent progress while acknowledging persistent hurdles. Addressing these challenges is vital for advancing the field as spam complexity increases.

Keywords: Spam, Social Networks, Communication, Summary of current spam identification methods, Combating Spam.

1. INTRODUCTION

The field of computer science basically describes unsolicited or unwanted messages that are generally transmitted electronically in bulk as spam or electronic spam, it is often sent through email or other communication channels such as social messaging apps, online forums, social media, blogs, newsgroups, web search, mobile phones and so on. These spam messages are usually for commercial purposes and often contain advertisements, fraudulent schemes, scams, phishing, spreading of malware etc (Wikipedia, 2023).

Spamming is considered a nuisance and can have numerous negative consequences and can lead to an overwhelming influx of unwanted messages, clod up inboxes, wasting of time and resources and it can also potentially expose an individual and organization to phishing attempts, malware and some other online threats (Sultana, et al., 2022).

Spam has essentially taken over just about every electronic platform across every media, and it has spread throughout the following media.

- *Email Spam:* This is also called junk email; it simply means unwanted or unsolicited email communications that are delivered in mass to a wide range of recipients without their consent. Email spam consumes a large amount of recipient network bandwidth and wastes recipient's time while dealing with them (Weisen, 2022).
- *Spam through Internet-based phone calls:* This is also called SPI, it is the practice of spamming over voice over internet protocol (VOIP), this is easier because of the ability to manage self-asserted personas while not having a subscription contract or identification documentation, internet telephony attracts a lot of telemarketers and fraudsters who make unwanted annoyance calls, and can rejoin a network with their numerous identity if they are found out (Javed, et al., 2021).
- *Spam in instant messaging (IM):* This use Instant Messenger i.e Whatsapp, Facebook Messenger, WeChat, Telegram and so on for spam distribution, this may, however, be subtle compared to email spamming, but it has the tendency of annoying user by including unrequested advertisement from advertisers and other sources.
- *Mobile phones Spam:* This distributes spam through Short Messaging Services (SMS) which occasionally has the ability to deceive users into signing up for false subscriptions and other scams by manipulating them. Customers are often unintentionally interrupted and become irritated by these unsolicited text messages.

- *Video sites Spam:* Spam on video streaming sites like YouTube typically consists of comments with links to dating websites, pornographic websites, or other irrelevant videos. Some of these comments are generated automatically by bots to spam the comment section.
- *Search engine spam:* web search spam also know as Spamdexing is a deliberate attempt to manipulate search engine relevancy and ranking in order to favour one or more web pages or websites, which could have a negative impact on the quality of the search engine's results. (Shahzad, et al., 2020).
- *Spam on blog and wiki:* This is used to describe comments that are off-topic and frequently contain external URL connections to websites that are either commercial, phishing or pornographic. Similar comments of this nature can also be seen in Wikis and other guestbooks that permit comments from any user.
- *Social networking spam:* Spamming can occur on social networking sites when they are used for a variety of objectives, including chatting, making friends or followers, consumer involvement, business purposes, product reviews and promotion, news, and online games. Users occasionally receive messages with spam links, which can reduce their online interaction time and lower the quality of the information that is made available. Also some automated accounts or bot often spread fake news, bogus review, spread rumours and spam massively to their target users. (Sanjeev, et al., 2021)

Customer ratings and reviews on the internet have developed into a reliable method to gauge public opinion about supplied goods and services in recent years as businesses have started to offer goods and services online. Therefore, because they can directly affect their operations, manufacturers and sellers place great value on online customer reviews. Spam reviews, on the other hand, are increasingly being written by users to either promote or denigrate certain goods or services. The validity and dependability of customer review-based business processes have come under scrutiny because of this practice, also known as review spamming. Despite the fact that academics have given the spam review detection (SRD) issue a lot of attention, the majority of SRD studies so far have used datasets in English, Arabic, Chinese, French, Spanish Persian, and other widely used languages. (Hussain, et al., 2021).

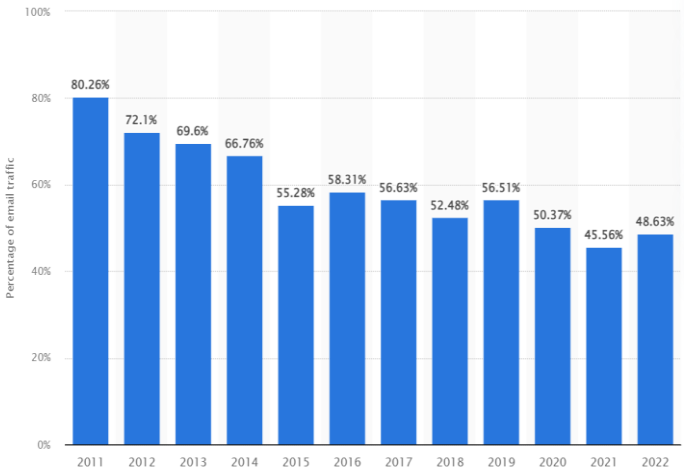


Figure 1: Global Spam email percentage. Source (Statista, 2022)

2. BACKGROUND

The term "Spam" originated from a Monty Python sketch that first hit the television screen in 1970, in which it mocked the prevalence of Spam luncheon meat. It later became associated with unsolicited junk emails. The word "Spam" does not stand for anything in the context of email spam, it simply refers to unwanted or any unsolicited commercial messages or broadcast messages which contain spam, they are a common problem and are sent to many recipients, outnumbering even the legitimate emails being sent across the globe (Leslie, 2022).

According to (ICTEA, 2023), it was reiterated that Initially, since spam was mostly limited to email and did not pose a serious threat other than to annoy recipients and waste their time, no serious measures were taken to combat it. However, as time went on and internet technologies advanced, people started using spam for marketing, and the majority of them have converted it into an illegal means of getting money. In some cases, spam has even been utilised for identity theft and other fraudulent crimes. Since its inception, email spam has changed substantially, with a growing proportion of emails now being spam that are sent via networks of infected computers, botnets, and a variety of other methods.

Since recipients bear the majority of the cost associated with spam, it can be seen as a form of advertisement where the recipient pays the "postage." Spam emails frequently include malicious software in the form of scripts or file attachments. Users may visit phishing or malware-hosting websites by clicking links in spam emails. Some emails also impersonate legitimate organizations to promote their content, These

deceptive advertisements entice the recipient to participate in contests or lotteries, tricking them into revealing sensitive information such as credit card details. Spammers typically gather email addresses from various sources like chat rooms, bulletin boards, customer lists, and forums, it could also be through a fake job opening where people are required to submit their details including phone numbers and email through a form. which are then harvested or shared with other spammers for malicious purposes (ICTEA, 2023).

Once the threat of email spam was recognized, efforts were made to implement various techniques to filter and combat it. A comprehensive discussion was provided by (Antonov, et al., 2021) on different email spam filtering techniques in their papers and delve into recent developments in email filtering and categorize the techniques based on several machine learning algorithms that were chosen for comparison, a natural language processing methodology was used to examine an email's text in order to identify spam. Decision Tree, Naive Bayes, SVM, Logistic Regression, and Random Forest are some of the algorithms used. By integrating algorithms of filtering techniques, we can use the outcome to produce a spam detection classifier that is more intelligent. (Antonov, et al., 2021).

The subsequent spam wave specifically targeted web search engines, The main purpose of search engines is to locate specific resources on the internet by using user-provided keywords or characters to conduct searches for specific items in databases (GCFGlobal, 2023). Search engine spamming, also known as spamdexing, is the practice of manipulating search engine results to seem as desired results for users. Search engines serves as a platform for showing results depending on user queries. Spamdexing is the deliberate manipulation of search engine indexes using a variety of techniques, such as the repetition of irrelevant terms, to alter the relevance or visibility of indexed resources in a way that is contrary to the indexing system's intended purpose. (Shahzad, et al., 2020).

In recent years, due to their rising popularity and the expansion of online social networks like Facebook, Twitter, Instagram, Tik-Tok, and others, individuals spend a lot of time on social media platforms. Unfortunately, this rise in usage has also attracted spammers, who are malicious individuals seeking potential victims, due to the freedom of content creation over social media, Malicious spammers take use of this chance to publish bookmarking links to business-related websites with the intention of phishing, infecting other users with malware, and stealing their personal information. When spammers acquire access to a user's profile, they can swiftly exploit that as a way to mine more data and use that to access their other accounts, sometimes asking for money on the user's behalf. Getting access to a cooperate social media account is a fantastic illustration of this. Spammers now have a new channel through which to commit cybercrime thanks to social networks, the emergence of these platforms has led to a significant increase in financial cybercrimes and frauds. False advertisements on social networks often deceive users, resulting in unfair trading practices. Spam does not discriminate based on age and can expose both young and adult users to inappropriate content, hindering moral development, especially in children. Additionally, spam contributes to piracy among other issues. However, spam is a major problem for the businesses that run social networking platforms. A social network's popularity and productivity depend on how many members are actively using it. However, consumers tend to avoid sites containing a lot of spam content because of the annoyance and disruption it causes. (Al-Zoubi, et al., 2019).

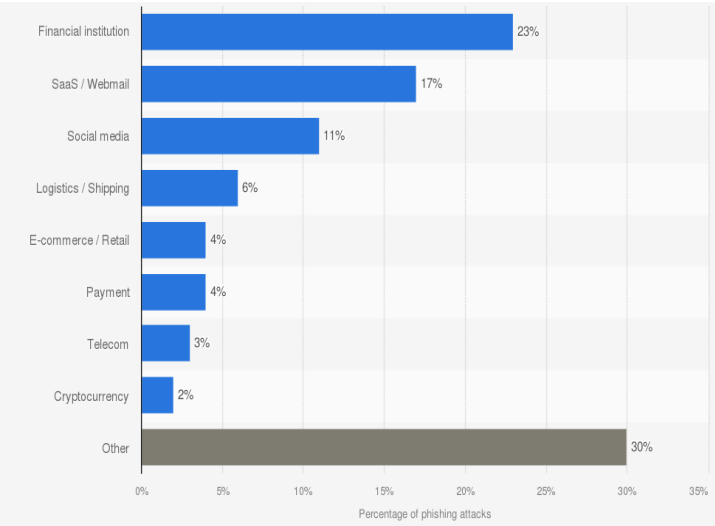


Figure 2: Online industries most targeted for Spams and Phishing. Source (Techopedia, 2023)

3. TYPE OF SOCIAL SPAM AND SPAMMING TECHNIQUES

In social spamming, fake accounts are essential since the goal is to appear credible by connecting to verified accounts, such as those of famous figures or celebrities, in the hopes of receiving reciprocal connections. When legitimate accounts accept these fake accounts as friends or followers, it lends credibility to the fakes and enables them to engage in spamming activities. Spammers may also hijack and gain control over a user's account, allowing them to disseminate false messages to the user's genuine followers. With the aid of these fake accounts, spammers can carry out various activities on social networks or applications, social spam can be broadly categorised into several types, including. (Tolentino, 2015)

1. *Bulk Messaging:* Messages containing identical or similar content can be rapidly distributed to a group of individuals. Additionally, multiple spam accounts can simultaneously share duplicate messages by utilizing bulk messaging, it is possible to artificially manipulate the popularity of a particular topic if enough people engage with it. For instance, in 2009, a spam website posing as a legitimate job opportunity at Google deceived users into believing its authenticity. Similarly, bulk messaging can be employed to disseminate malware or promote advertisements that redirect users to specific websites (Tolentino, 2015).

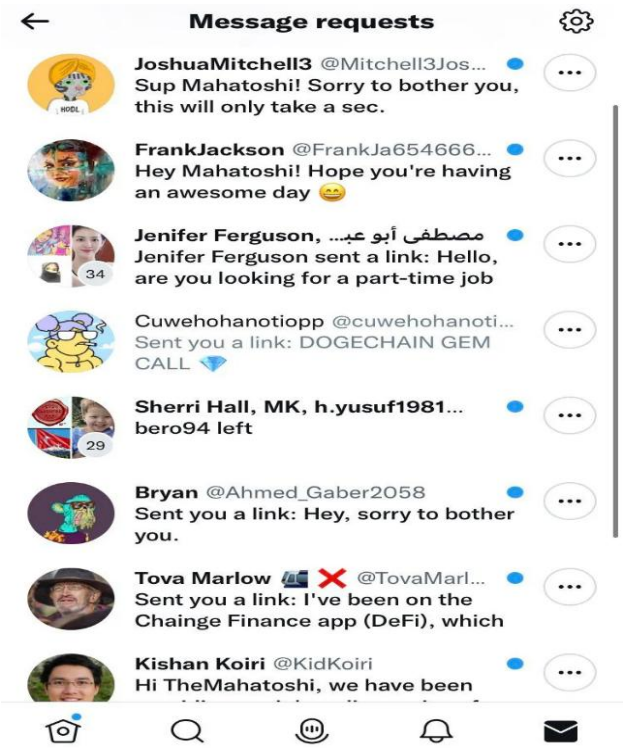


Figure 3: An example of bulk messages on Twitter. Source (Reddit, 2022)

2. *Malicious Link:* these are created with the intention of causing harm or deception to users or their devices. When clicked, these links can lead to various harmful activities such as downloading malware or stealing personal information, they are often spread through user-submitted comments and posts on platforms like YouTube, Facebook and Twitter. Additionally, fake social media accounts can also distribute these links through posts or messages.

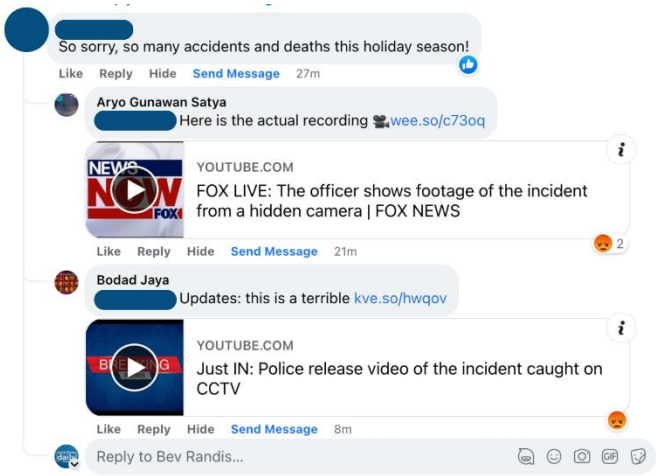


Figure 4: An example of a malicious link posted on social media. Source (KRTV, 2022)

3. *Fake Review:* This refer to reviews written by users who have never actually used the product or service being reviewed. In many cases, in order to strengthen the public perception of their goods or services, businesses or individuals may pay customers to leave favourable reviews. By utilizing fake accounts, these fabricated reviews can be easily posted under false identities, often in large quantities.

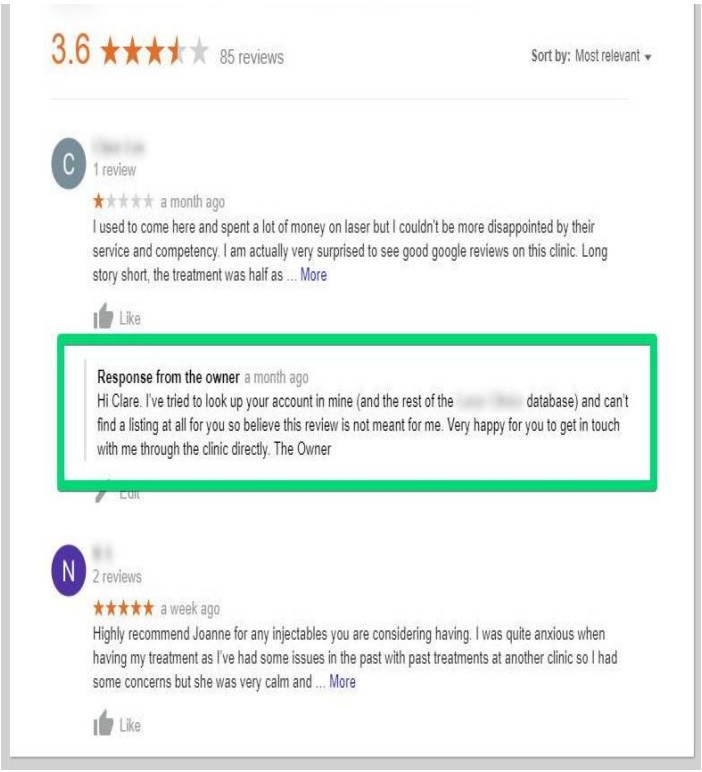


Figure 5: Fake review comment Source (DigitalMaas, 2018)

4. *Fake Profile:* To avoid detection and entice other trusting users to friend or follow them, spammers may construct phoney profiles that otherwise appear real. phoney profiles are occasionally established for amusement or to cause trouble, and sometimes spammer use it to spread misinformation or solicit for funds from some of their naïve followers.

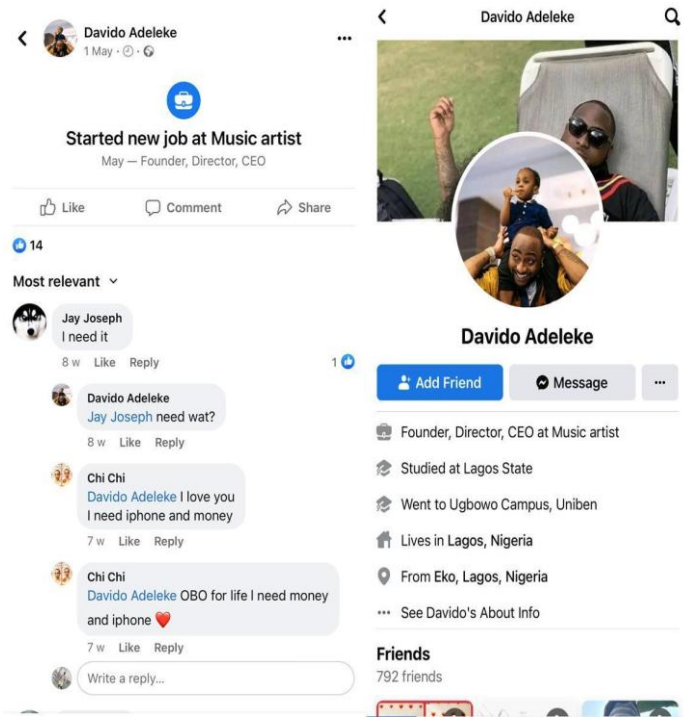


Figure 6: An example of a fake profile on Facebook. Source (Facebook Screenshots)

(Jain, et al., 2021) Consider the risks listed above to be conventional spam threats (threats that users have faced since the inception of social networks), while the following are attacks that employ cutting-edge methods to compromise users' accounts.

- *Cross-site scripting attack:* also known as XSS attack, is a common attack where malicious JavaScript is run on the victim's browser using a variety of tactics, after which links and buttons on social media websites can be used to fool users into clicking URLs that could lead to spyware or information theft.
- *Profile cloning attack:* Attackers utilise a user's cloned profile on social media networks to gain the trust of the user's friends, and they then fool them into disclosing private information that can be used for cyberbullying, cyberstalking, or extortion.
- *Hijacking:* Adversaries compromise or take control of user accounts to carry out online fraud, Weak passwords and lack of multifactor authentication make accounts vulnerable. Hijackers can send messages, share malicious links, or modify account information, damaging the user's reputation.
- *Inference attack:* By studying publicly accessible data on social media outlets, including friend lists and network architecture, attackers might deduce private information that can reveal organisational secrets or user personal information.
- *Sybil attack:* In an attack known as Sybil, a node in a network makes several identity claims, particularly impacting social networking platforms with a large user base, fake identities are used to spread misinformation, malware, or manipulate online surveys.
- *Clickjacking:* Attackers deceive users by making them click on a different page than intended, exploiting browser vulnerabilities, this includes variations like likejacking, where users unknowingly click on the "like" button, and cursor-jacking, where the actual cursor is replaced with a custom image.
- *De-anonymization attack:* Users on social networking sites can hide their real identity using aliases, but attackers can link leaked information to uncover their true identity, techniques like tracking cookies and network topologies are used to de-anonymize users.
- *Cyber espionage:* Cyber espionage is the use of technological tools to obtain confidential data or intellectual property; it is frequently done for financial gain or as part of military operations. Attacks using social engineering on websites for social networking can be used to get important data.

4. DETECTION TECHNIQUES

4.1 Review of conventional methods

Social spam requires a distinct approach compared to other spam types like email and web spam. Its unique characteristics necessitate a different approach to tackling it. The researchers outline a social spam model that can encompass different types of social spam. Additionally, (Balogun, et al., 2017) categorize various strategies for combating social spam into three broad categories.

- *Preventive Base:* This approach aims at preventing spam content from affecting social tagging systems which can be done by implementing restrictions on certain access points, such as using CAPTCHA tests or imposing usage limits. i.e platforms like Flickr introduced a limit of 75 tags per photo to control tagging spam. These measures are to ensure it is more difficult for automated systems or excessive tagging to contribute to social tagging systems, ensuring a better user experience and more accurate content organization.
- *Detection-based approaches*: involve the identification of potential spam through manual or automated means, then machine learning techniques, like text classification, or statistical analysis methods, such as link analysis, are utilized to identify likely spam content, once it is identified, the spam content is either deleted or visibly marked as hidden to the user.
- *Demolition Based:* This approach aims to minimize the visibility of potentially spammy content, one method used is rank-based algorithms that generate an ordered list of a system's tags or users based on their trust score.

It has been argued that the earlier methods for detecting social spam lack specificity when it comes to topic extraction because they primarily rely on user data and metadata to passively extract features. Additionally, further analysis is made more difficult by the difficulty of properly understanding the historical significance of social written content. Existing methods for detecting social spam focus on textual analysis using probabilistic generative models like Latent Dirichlet Allocation, also known as LDA, and its variants (e.g., Labelled LDA, PhraseLDA). These models, however, have a number of drawbacks:

i. They struggle to understand higher-level concepts or domains

ii. They miss the semantic connections between keywords in the text

iii. They are insufficient for extracting subjects from brief textual content, like tweets, or for drawing conclusions from isolated texts (Abu-Salih, et al., 2022).

4.2 Latest Developments in Spam Detection

The continuous development of social networks and their heightened security measures have compelled spammers to adjust their tactics accordingly. As a result, some of the most intricate forms of spam have emerged in recent times. (Jain, et al., 2021) provide an overview of certain contemporary spamming techniques and briefly touch upon anti-spamming methods. In this discussion, we delve into additional proposals, focusing particularly on the variety of these methods. Based on the particular domains in which they are used, we classify anti-spam strategies as follow.

4.2. 1 Email-Spam

Although networking sites and social media spam are the main emphasis of this article, we also briefly examine several methods for filtering email spam. (Zavrak & Yilmaz, 2022) presents an innovative approach to detecting email spam using a synthesis of attention mechanisms, gated recurrent units, and convolutional neural networks. The network specifically concentrates on pertinent portions of the email text during the training phase. The application of convolution layers in order for gathering more significant, abstract, and generalizable information through hierarchical representation is a significant contribution of this study. Additionally, the study uses cross-dataset evaluation, enabling the creation of separate performance findings from the model's training dataset.

By utilising temporal convolutions, which offer flexibility in receptive field sizes, the strategy outperforms existing attention-based methods, as shown by the results of the cross-dataset evaluation. This strategy outperforms state-of-the-art models in comparison, which validates the usefulness of this technique.

Hossain, et al., developed a model that distinguishes between spam and ham emails. To detect values that are outside of a particular range, it employed DBSCAN (Density-based clustering algorithm) and Isolation Forest algorithms. They used Chi-Square feature selection, Heatmap and Recursive Feature Elimination strategies to choose useful features, to enable comparative examination, the model was developed using Machine Learning (ML) as well as deep learning techniques..

The K-Nearest Neighbour (KNN), Multinomial Naive Bayes (MNB), Gradient Boosting (GB) and Random Forest (RF) algorithms were used as part of an ensemble method in the machine learning implementation. Artificial Neural Networks (ANN), gradient descent (GD) and Recurrent neural networks (RNN) were used for the deep learning implementation. The output of several classifiers was combined using an ensemble method, which increased prediction accuracy in comparison to using only one classifier.

This model produced a precision of 100%, AUC=100, RMSE error = zero and MSE error = zero, in the ML implementation using an email spam base dataset that was obtained from the UCI machine learning repository. A loss value of 0.0165 and an accuracy of 99% were achieved with Deep Learning implementation (Hossain, et al., 2021).

Table 1: Email Spam

Paper	Task	Methods	Dataset used	Outcomes
(Hossain, et al, 2021)	Filtering	Machine Learning & Deep Learning	UCI Machine Learning Repository	99% accuracy

(Zavrak & Yilmaz, 2022)	Classification	Convolutional Neural Network	Cross dataset	Outperform existing attention-based methods

4.2. 2 Blog-Spam

A popular forum for people to communicate, share information, and express their feelings is blogging. As blogs have become more popular, they are also being used for advertising and to draw visitors from blog search engines. However, this has led to the emergence of spam blogs, many existing techniques for spam blog detection primarily rely on content-based approaches, which may be less effective due to the dynamic nature of blogs. (Li, et al., 2019) Propose a study that specifically addresses the detection of comment spam, it entails examining spammers' actions and spam's content.. Through this analysis, two types of features were identified as effective for providing a more accurate description of spammer characteristics, to construct the comment spam detector, a gradient boosting tree algorithm was employed using these extracted features. The suggested methodology was assessed using a blog spam dataset from previous research, and the outcomes demonstrate that our strategy exceeds the earlier technique with regard to of detection accuracy. Additionally, the CPU time was measured to demonstrate that both training and testing processes are executed efficiently within a short timeframe (Li, et al., 2019).

By using a Cascaded Ensemble ML Model and comment dataset from popular music videos by LMFAO, Katy Perry, Shakira, Psy, and Eminem (Hayoung, 2021) proposed a novel method to identify spam comments in YouTube comment data. Using six different machine learning techniques (Bernoulli Naive Bayes, Logistic regression, Support vector machine with Gaussian kernel, Random Forest, Decision tree and Support vector machine with linear kernel) as well as two ensemble models (Ensemble with soft and hard voting) applied to comment data, the researcher's study reviewed previous studies on YouTube spam comment screening.

The experimental findings showed that, in four of the five evaluation metrics, the suggested ESM-S model performed better than alternative models. He also expanded the applicability of the ensemble model to movies from numerous categories, in contrast to earlier studies that only used a single model for detection. In both datasets, the ESM-S model consistently outperformed other models in terms of Matthews correlation coefficient (MCC), F1-score and Accuracy (Acc), Spam Coverage (SC) was a strong suit for the Artificial Neural Network (ANN) model, while Balanced Hit Rate (BH) was a strong suit for the Naive Bayes-Bernoulli (NB-B) model.

Additionally, the experimental findings showed that the dataset with 5,000 (five thousand) comments performed worse than the dataset with 1,000 (a thousand) spam comments and 1,000 (a thousand) regular comments, most likely because of missing values and possible outliers.

Webedia developed and implemented a machine-learning algorithm to detect spam blogs on Overblog, a blogging platform. The increase in popularity of Overblog attracted spammers who posted malicious content, negatively impacting the user experience and platform quality. To address this issue, a training dataset was manually created by categorizing posts as spam or legitimate, A Random Forest model was trained using features including the number of linkages and NLP methods, which maximised accuracy and minimised false positives, this model was deployed as a REST API, allowing the Overblog team to obtain spam predictions, the system could adapt to new spammers through a feedback mechanism, continuously improving detection. This approach successfully cleaned the database and reduced the influx of spam, resulting in improved platform traffic without deleting legitimate users' blogs. The machine learning system provides flexibility and confidence in combating spammers (NGUYEN-KHOA-MAN, 2022).

Table 2: Blog Spam

4.2. 3 Microblog Spam

A microblog refers to short content intended for swift interactions with the audience. Microblogging merges features of content creation and instant messaging, allowing its user to post short messages with an online audience to enhance engagement. Microblogging is very common on well-known social media sites like Facebook, Pinterest, Instagram and Twitter (SproutSocial, 2023).

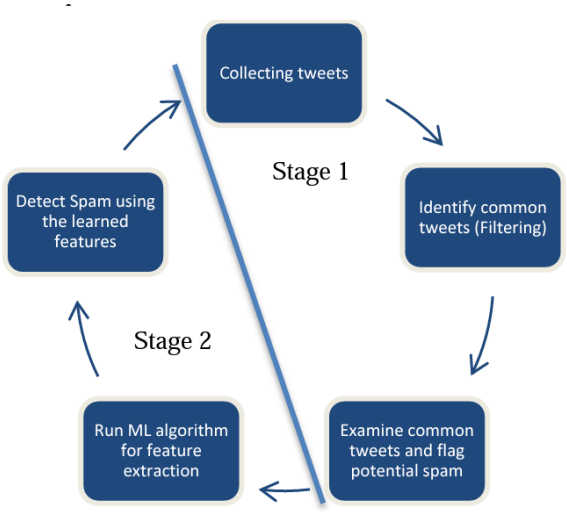
In a paper presented by Kabakus and Kara using Twitter as a case study, the most widely used microblogging platform, which allows users to share short status messages known as tweets, this popularity, coupled with Twitter's robust API for programmatically accessing and manipulating Twitter data, attracts both legitimate users and spammers. Due to the unique characteristics of Twitter, Using standard spam detection techniques to find spam on this platform is not a direct solution. The paper proposes TwitterSpamDetector, a specialised spam detection technology created especially for Twitter, to handle this issue. TwitterSpamDetector leverages Twitter-specific features to identify spam accounts and content. A set of data of 77,033 tweets published by 50,490 people and obtained using Twitter's API is used to train the framework. The selected Twitter features are used in conjunction with the Naive Bayes algorithm to train the TwitterSpamDetector, enabling it to effectively distinguish spammers from legitimate users. Evaluation results reveal that TwitterSpamDetector achieves an accuracy of 0.943 and a sensitivity of 0.913, demonstrating its effectiveness in detecting spam on Twitter (Kabakus & Kara, 2019).

According to (Binsaeed, et al., 2020), The protection of sensitive data, IT infrastructure, and financial assets from

unauthorised actions like identity theft, worm downloads, and extortion is of the utmost significance. The identification of hostile activity on the internet continues to be a crucial issue that requires appropriate methods. Malicious individuals now have a fresh and promising channel to carry out their actions, from sending out simple spam messages to taking complete control of their victims' computers, thanks to the rise of social networking platforms like Twitter. The abundance of irrelevant spam and promotional tweets in popular hashtags shows that Twitter's current spam detection algorithms are inadequate. It is clear that there is opportunity for improvement in the framework used for spam detection now. In order to identify spam in Twitter microblogging, this study presents a novel method that makes use of machine learning (ML) tools and domain popularity services. The suggested strategy is divided into two phases:

1. Periodic collection of tweets/posts and filtering based on a predetermined threshold for frequency within a specified period, identifying common tweets. The corresponding URL domain of these frequent tweets is then checked against Alexa's leading one million globally accessed websites as part of a further analysis. A tweet is marked as potentially spam if it is popular on Twitter yet fails to appear among the top a million domains.

2. In the second stage, features that help in the real-time detection and prevention of spam clusters are extracted from the flagged tweets using Machine Learning algorithms. Three well-known classification models (J48, random forest, and Naive Bayes) were used to assess how well the suggested technique performed.



Paper	Task	Methods	Dataset used	Outcomes
(Li, et al, 2019)	Detection	Gradient boosting tree algorithms	Dataset from the previous search	Outperform previous methods
(Hayoung, 2021)	Classification	Ensemble with hard and soft voting	Youtube video comment data	Higher F1 and MCC score than Naïve Bayes-Bernoulli (NB-B) model
(NGUYEN-KHOA-MAN, 2022)	Detection	Machine learning	A manually created training dataset	Successfully clean the database

Figure 7: Model outline. Source (Binsaeed, et al., 2020)

Results across all classifiers showed that the proposed strategy was effective in terms of a variety of performance measures, including F1-score, precision, as well as accuracy and sensitivity, under various test situations. The final accuracy results are summarized in table 3 below.

The research paper by (Kardaş, et al., 2021) discussed an important feature of spamming activities, online social networks (OSNs) have gained immense popularity, making them attractive platforms for spammers. These spammers exploit OSNs to easily disseminate malicious content and promote phishing scams. Consequently, the identification and filtering of spam tweets have become crucial for both OSNs and users. However, the sheer volume of posts makes it increasingly challenging to detect and eliminate spam tweets. Motivated by this scenario, this research paper proposes an approach that utilizes ML and effective preprocessing methods for detecting spam tweets. The paper highlights the benefits of preprocessing

and identifies the most effective techniques through comparative analysis using the UtkML Twitter spam dataset. Once the most effective methods are determined, they are combined to optimize the accuracy of spam tweet detection. Four different machine learning algorithms— Support Vector Machine, Logistic Regression, Naive Bayes Classifier and Neural Network—are used to evaluate the answer. The SVM Classifier achieves a 93.02% accuracy rate. The experimental findings show that their method considerably improves spam tweet classification performance.

Table 3: Microblog Spam

Paper	Task	Methods	Dataset used	Outcomes
(Kabakus & Kara, 2019)	Detection	Naïve Bayes algorithm	Tweets from Twitter API	Up to 94% accuracy
(Binsaeed, et al, 2020)	Classification	Random forest, J48, and Naïve Bayes	Twitter API	Accuracy: Random Forest = 92% J48 = 95% Naïve Bayes = 84%
(Kardaş, et al, 2021)	Filtering and classification	Neural Network, Logistic Regression, Support Vector Machine	UtkML Twitter spam dataset	93.02% accuracy achieved

4.2. 4 Social Network Spam

Establishing an engaged social media audience can be challenging, requiring significant time and effort to create content that resonates with your target audience. Striking the right balance between valuable content that drives growth and spammy content that repels potential customers is crucial. Busy Business-to-Business (B2B) decision-makers don't have the luxury of consuming repetitive content that doesn't meet their needs. Instead, they seek relevant, data-driven content throughout their journey. Ensuring that your social media posts consistently hit the mark is essential. Spam on social media extends beyond content creation. Some B2B brands produce excellent, actionable content, but their accounts may be labeled as spammy due to an influx of comments, some of which promote scams. This situation becomes even more problematic for unverified business accounts. According to (Orika, 2022) in a survey conducted, Facebook is the most spammy social media. See fig below

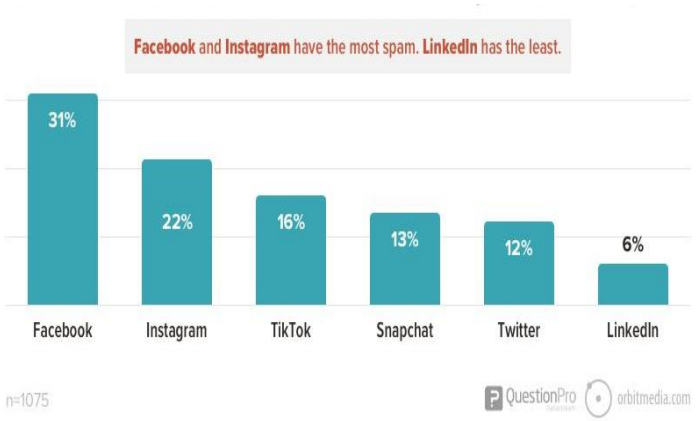


Figure 8: Ranking of most spammy social media. Source (Orika, 2022)

As they connect a sizable user base of over two billion accounts, Twitter, Facebook, and LinkedIn, among others, are largely acknowledged as the most important and potent platforms on the web. Through these online social networks (OSNs), relationships between family members and friends have increased to previously unheard-of heights. Several factors contribute to the growing popularity of these OSNs. First of all, they offer an accurate portrayal of actual social interactions between people. Their simple to operate web-based platforms also make it simple and quick to post many kinds of user-generated content. From a financial perspective, these platforms give businesses and governments numerous chances to examine consumer behaviour, receive feedback, and sell their goods successfully. All these appealing features have made OSNs the preferred choice for individuals across the Internet. As presented by (Al-Zoubi, et al., 2019) in their article which focus

on the issue of spam profiles in online social networks, particularly in platforms like Twitter, which pose a significant security threat on the Internet. If these spam profiles continue to produce unwanted adverts, criminals may use them for a variety of harmful objectives. By examining the nature and traits of spam characteristics using readily accessible, language-independent criteria, this paper aims to improve spam detection. Four datasets from various language contexts (Korean, Arabic, Spanish and English) were collected to assess the efficacy of these variables in spam detection, and putting them together yields a fifth dataset. Five well-known classifiers—Multilayer Perceptron (MLP), Decision Tree (DT) (J48), k-Nearest Neighbours (k-NN), Naive Bayes (NB), and Random Forest (RF), classifiers—are employed in the tests to identify spam. Additionally, ReliefF, the Information Gain, Correlation, Chi-square, and Significance filter-based feature selection approaches are used. The results show that each classifier performs differently across the datasets, but that utilising feature selection generally improves classification outcomes. Additionally, thorough comparison and analysis of selected features were conducted on two levels; the first level compared the significance of selected features among feature selection methods, and the second level looked at the relationships and significance of chosen elements within all datasets. The conclusions of this article improve understanding of social spam and aid in the development of detection methods by taking into account the significant features derived from feature selection methods. After the filtering, the result for each language is summarized below:

- ***Arabic Dataset:***
  - k-NN with ReliefF method and top 20 selected features achieves 98.4% accuracy.
  - ReliefF shows the highest precision Recall, F-measure, and AUC with 99.1%, 98.6%, and 99% respectively.
  - Using the top 20 and 15 features, respectively, the Chi-square and Significance techniques obtain 98.5% precision.
- ***English Dataset:***
  - k-NN with ReliefF method and top 25 selected features achieves 97.7% accuracy.
  - ReliefF performs well in Recall, F-measure, and AUC.
  - The approaches InfoGain, Chi-square, and Significance have the highest levels of precision.
- ***Korean Dataset:***
  - J48 classifier achieves 92.4% accuracy using InfoGai n's top 25 features.
  - Recall and F-measure metrics are where J48 excels.
  - AUC outcomes are constant across all filters.
- ***Spanish Dataset:***
  - With the top 25 features chosen by both ReliefF and InfoGain, the NB classifier achieves 90.9% accuracy.
  - NB achieves the best F-measure and Recall results.
  - Improved accuracy compared to the original dataset.
- ***Multilingual Dataset:***
  - RF classifier achieves 95.23% accuracy using Significance's top 25 attributes.
  - Shows slight improvement in Precision, Recall, and F-measure.

Chaudhry, et al. reiterate that the social networks are susceptible to the influence of spammers, and significant efforts have been undertaken to recognise and resolve this problem. Support Vector Machine (SVM) was used by the researchers as a classification method for detecting spam in social networks. The effectiveness of their suggested approach is assessed using a number of different factors. In order to evaluate the success of their suggested work, the methodology used to detect spam in a social network, they also conducted a comparative analysis between it and existing methodologies. Training and testing are the two stages of the spam detection process. The characteristics of the dataset are calculated and given weighted values during the training phase. These feature sets are then used to train the Support Vector Machine (SVM), which then stores the information in its database. The test data needed for spam detection is loaded as we move on to the testing phase. Features are weighted and calculated. The features from the test are then contrasted with the database's features. The data is classified as spam if the features match; else, it is thought to be a regular message. Parameters such as F-measure, recall and the precision were used in the study, and the resulting values for each of them were 84.01, 80.99, and 87.38. Finally, performance metrics are evaluated in order to measure the efficacy of the created spam detection model (Chaudhry, et al., 2020).

In their report paper Koggalahewa, et al., reported that detecting spammers in online social networks (OSNs) is a highly challenging task. The majority of current methodologies rely on supervised classification techniques, yet these techniques have drawbacks such unbalanced datasets, data labelling, data falsification and spam drift. The reliability of spam recognition classifiers is significantly impacted by these constraints. In their study, they provide an unsupervised method for separating spammers from legitimate members in social networks by relying on peer acceptance. Their method determines a user's peer acceptance based on shared interests in a variety of topics. A two-stage unsupervised method for spam detection is presented in this paper. In the initial phase, clustering algorithms were utilised to divide consumers into two distinct groups: "Focused" and "Diverse." This classification assesses whether individuals have a focused or diverse interest across a variety of topics based on the distribution of their information interests. In the second stage, users are assessed using our suggested peer-acceptance-criterion, which looks for possible spammers by analysing shared interests across a variety of areas.

To assess the effectiveness of this strategies, they utilize publicly available datasets, namely TheSocialHoneypot, HSpam14, and TheFakeProject datasets. Latent Dirichlet Allocation(LDA) was used for clustering and Peer acceptance based on binomial distribution was used for classification, this paper's major contribution is the development of a completely unsupervised spammer detection method that does not require labelled training datasets. Their solution, which achieves a 96.9% accuracy rate for spam identification, is an effective alternative even though it may not be as accurate as supervised classification-based methods. (Koggalahewa, et al., 2021).

Shen, et al. present an article that describes a novel approach to identifying spam in social media photographs by combining deep learning with frequency domain pre-processing. Few studies have focused on identifying the presence of spam in images, whereas typical methods tend to concentrate on detecting spam in links and texts. To fill this gap, the suggested method combines deep neural networks with frequency domain pre-processing. In order to train the detection model, the method requires gathering a dataset of photos with embedded spam and integrating it with the DIV2K2017 dataset. A pre-processing module is created in accordance with the studies that identify the precise spam components that are present in the photos. In the pre-processing module, low-frequency domain regions with lower spam levels are discarded using Haar wavelet transform analysis. In order to maximise the acquisition of high-frequency features from three different regions, a feature extraction module is also created employing unique convolutional layers. The final classification outcome is then produced by concatenating the extracted high-frequency features along the channel dimension. Numerous experimental findings show that the majority of spam components are found in photos as high-frequency information bits. When utilizing the high-frequency components as inputs to the model, the detection accuracy reaches a high of 86%. However, when using the entire image as input, the detection accuracy significantly decreases to only 74.5%, modern detection models can't compete with the model's efficiency and accuracy in terms of detection. (Shen, et al., 2022).

Two methods are used to detect spam in online social networks (OSNs): machine learning (ML) and expert-based detection. Expert-based identification relies on human expertise and involves a manual and time-consuming process, making ML-based spam detection more preferable in OSNs. The difficulty of spotting spam on social networks is influenced by a number of factors, and the dispersion of spam and genuine content provides spammers a leg up in infiltrating our devices. To handle the issue of unbalanced data and produce reliable assessments, ML algorithms like K-Nearest Neighbour (KNN), Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), Voting Classifier (VC), XGB and Support Vector Machine (SVM) are frequently employed for spam identification. Text is vectorized with the use of vectorizers, and the pertinent outcomes are saved. Compared to existing algorithms like DT, NB, SVC, ETC, KN, RF, XGB, and LR, experimental findings show that the proposed Voting Classifier (VC) gets a superior classification accuracy rate of 97.96%. The dataset used to train the models is made up of more than 5500 data messages that were gathered from the data science company "Kaggle". The results demonstrate that the methods they provided are successful in detecting both balanced and imbalanced datasets. (Sumathi & Raja , 2023).

Table 4: Social Networking Spam

Paper	Task	Methods	Dataset used	Outcomes
(Al-Zoubi, et al, 2019)	Filtering and Classification	Random Forest, Naïve Bayes, Decision Tree with Multilayer Perception Specifier (MLP)	Twitter REST API dataset from Tweets	Up to 97% in English Language dataset
(Chaudhry, et al, 2020)	Classification and detection	Machine Learning with Support Vector Machine (SVM)	Facebook Post	87.3% Precision
(Koggalahewa, et al, 2021)	Clustering and Detection	Unsupervised approach using Latent Dirichlet Allocation(LDA) for clustering and Peer Acceptance distribution	TheSocialHoneypot HSpam14 TheFakeProject	>96.9% Spam detection
(Shen, et al, 2022)	Detection	Deep Neural Network and Frequency Domain Pre-processing	Random images containing embedded spam and DIV2K2017 dataset	Upto 86% with high-frequency input data
(Sumathi & Raja , 2023)	Identification and Detection	Machine Learning using Logistic Regression (LR), K-Nearest Neighbor (KNN), Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), XGB, and Voting Classifier (VC)	Kaggle social media dataset	Close to 98% accuracy

4.2.5 Review Spam

Public opinions on various products or services are now primarily formed based on online reviews, this makes customer reviews crucial for manufacturers and sellers as they directly impact their businesses. However, the practise of "review spamming," in which phoney reviews are posted to either promote or denigrate particular goods or services, has raised some issues. Although communities and scholars have focused on the detection of spam reviews, there is still a need for research using practical significant review datasets to examine the pervasive effects of opinion spam. In a study by (N. Hussain, et al., 2020) which introduces two distinct methods for spam review detection:

1. The first is the Spam Review Detection Using Behavioural Method (SRD-BM), which uses thirteen behavioural features of spammers to calculate a review spam score and identify spammers and spam reviews. And

2. The second is the Spam Review Detection Using Linguistic Method (SRD-LM), which analyses the content of reviews and uses transformation, feature selection, and methods of classification to recognise spam reviews.

The findings of the assessments show that both models considerably improve the detection of spam reviews. The models were tested using an operational Amazon review dataset that included 26.7 million reviews and 15.4 million reviewers. For example, SRD-BM detects spam reviews with a precision of 93.1%, whereas SRD-LM does so with an accuracy of 88.5%. Notably, SRD-BM beats SRD-LM in accuracy since it makes use of a wide variety of behavioural characteristics of spammers, enabling in-depth research of spammer conduct. Both models outperform current methods in correctly identifying spam reviews.

Shahariar, et al. Explain the urgent need for a robust and reliable system to detect spam reviews in order to make trustworthy online purchases. The presence of options for posting reviews on many online platforms creates opportunities for fake paid or misleading reviews. This can confuse the general public and make it challenging for them to determine the authenticity of reviews. While supervised learning techniques have been extensively used in the detection of bogus reviews, they rely heavily on labelled data, which is often insufficient in the context of online reviews. Two datasets were utilized in the experiments: a labelled dataset known as the 'Ott dataset' and an unlabelled dataset consisting of real-life reviews from the overtly available Yelp dataset. Both the 'Ott dataset' and the 'Yelp Dataset' were employed in our experiments.

In that article, their focus is on detecting deceptive text reviews, to achieve this, They used both labelled and unlabeled data and proposed deep learning techniques such as Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN), as well as a variant of Recurrent Neural Network (RNN). For the purpose of identifying spam reviews, they also used well-known machine learning classifiers as Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Naive Bayes (NB). Finally, they contrast the effectiveness of deep learning models and conventional machine learning classifiers. (Shahariar, et al., 2022).

Table 5: Review Spam

Paper	Task	Methods	Dataset used	Outcomes
(N. Hussain, et al, 2020)	Detection	Behavioural Method and Linguistic Method	Amazon Website review dataset	>93% for the behavioural method >88% for the linguistic method
(Shahariar, et al, 2022)	Classification	Deep Learning i.e CNN, LSTM, MLP	Ott dataset and Yelp Dataset	CNN: >95% LSTM: >96% MLP: >93

4.2. 6 Location Spam

Location-based social networks (LBSNs), which give users a place for communicating and exchanging information depending on their geographic locations, have become an essential part of our daily lives. The openness of LBSNs, however, also leaves them exposed to malevolent users who spread false information in an effort to influence users' choices in urban computing environments. To guarantee the accuracy of the data and improve the user experience, A system called DeepScan was created by (Gong, et al., 2018) to find fraudulent accounts in LBSNs. In contrast to current methods, DeepScan makes use of deep learning techniques, specifically the long short-term memory (LSTM) neural network, to analyse users' dynamic behaviour over time. DeepScan uses a supervised machine learning model for detection by combining recently introduced time series features with traditional features extracted from users' activities. They assess its performance using genuine traces obtained from the well-known LBSN Dianping. The outcomes show that DeepScan has exceptional prediction accuracy, earning an extraordinary F1-score of 0.964. Additionally, their findings show how important time series features are to the detection system's efficiency.

In the journal of He, et al., the authors investigate this using dating apps, dating apps have experienced a surge in popularity over the past decade, revolutionizing the process of finding partners compared to traditional offline methods. However, this

increased usage also makes dating apps susceptible to malicious activities. The research focuses in identifying fraudulent users on dating applications. Previous approaches had limited detection performance because they failed to take into account the important signals that were concealed inside the textual records of user interactions, particularly the relationship between temporal-spatial activities and the content of the text. They suggest DatingSec, a cutting-edge technique for identifying rogue users in dating applications, to overcome this drawback. In order to accurately capture the complex interaction between users' temporal-spatial activities and the textual information they produce, DatingSec which uses long short-term memory neural networks (LSTM) and an attentive module. They use a real-world dataset gathered from Momo, a well-known dating app with over 180 million members, to assess the efficacy of DatingSec. Experimental results show that DatingSec outperforms cutting-edge techniques, with an AUC of 0.940 and an excellent F1-score of 0.857 (He, et al., 2021).

Caha & Kovar'k presented a spam filter that employs IP geolocation to pinpoint the location of the email sender. The filter was integrated as an extensions for the SpamAssassin spam filtering programme. Users of this plugin can give specific nations renowned for sending spam a penalty score. The suggested spam filter, known as Geolock, was integrated with the "IP2Location DB5.LITE geolocation database" and is freely accessible on GitHub. 1500 emails—1200 spam and 300 ham—from a bespoke dataset were utilised to assess the effectiveness of the spam filter. The filter's Matthews correlation coefficient, which was calculated to be 0.222, shows that it contributes to accurate spam filtering. The results also showed that the filter was successful in correctly recognising spam emails, with a precision value of 0.992 and a very high specificity value of 0.993. (Caha & Kovaˇr´ık, 2022).

Table 6: Location Spam

Paper	Task	Methods	Dataset used	Outcomes
(Gong, et al, 2018)	Analyze, Classification and Detection	Deep Learning	Dianping (LBSN)	>96 accuracy
(He, et al, 2021)	Detection	LSTM based on Neural Network	Momo dating app dataset	>85% accuracy
(Caha & Kovaˇr´ık, 2022)	Filtering and classification	Correlation Coefficient	Authors Email dataset	>99% precision

4.2. 7 Comment-Spam

Every part of our internet presence has been compromised with spam, making its way into various facets of our digital lives. One area where it is particularly prevalent is in the provision of free comment sections found on social media platforms, such as YouTube and news websites. These platforms have unfortunately become prime targets for spammers, who misuse these features to a significant extent. (Samsudin, et al., 2019) proposed using data mining to filter spam comments on YouTube forums., they noted that YouTube has gained immense popularity as a social media platform, but unfortunately, this popularity has also attracted spammers who distribute spam through YouTube comments. This is a major concern since spam can potentially lead to phishing attacks, targeting any user who clicks on malicious links, to address this issue, it is crucial to analyze and detect the specific characteristics of spam through classification techniques. As a result, their analysis proposes an improved feature set for accurately identifying YouTube spam. A thorough YouTube spam detection framework with five stages—data collection, pre-processing, feature selection and extraction, classification, and detection—was created in order to carry out the trials. This study uses two separate data mining techniques to introduce and validate each stage of the framework. Logistic Regression techniques and Naive Bayes were used to analyse data from the YouTube Spam dataset to create the features, which were then tested using Weka and RapidMiner.

The analysis revealed that thirteen features exhibited high accuracy when tested on both Weka and RapidMiner, thus making them suitable for the subsequent experiments in this research. In Weka, the results obtained from Naïve Bayes and Logistic Regression were slightly higher, with accuracy rates of 87.21% and 85.29% respectively. On the other hand, in RapidMiner, there was a slight difference in accuracy between Naïve Bayes and Logistic Regression, with rates of 80.41% and 80.88% respectively. However, Naïve Bayes demonstrated higher precision compared to Logistic Regression.

According to a paper by (Abinaya, et al., 2020) which claimed that spammers had gotten better at utilising a variety of tactics to persuade consumers to click on harmful links. They frequently use the technique of creating spam in the social media networks' comment areas. They used YouTube comments as their dataset for the study and carried out a detection analysis that targeted spam YouTube comments. Utilising technologies like Google Safe Browsing, which assist in identifying and eliminating irrelevant spam on YouTube, is one method now used to battle spammers. However, while these tools help to prevent hazardous links, they fall short of offering real-time security for users. As a result, a wide range of various strategies have been investigated to produce a spam-free environment. Some of these strategies are based on user-driven choices, while others are based on YouTube content. We used four different machine learning algorithms—Ada Boost Classifier, Logistic Regression, Support Vector Machine, Decision Trees Classifier, and Random Forest — to

assess the efficacy of our strategy. Among these algorithms, Logistic Regression outperformed the present solution by about 18%, achieving a remarkable accuracy rate of 95.40%.

Table 7: Comment Spam

Paper	Task	Methods	Dataset used	Outcomes
(Samsudin, et al, 2019)	Extraction, classification and detection	Naïve Bayes and Logistic Regression	Youtube comment dataset	>87% with Naïve Bayes. >85% with Logistic Regression
(Abinaya, et al, 2020)	classification	Various Machine Learning algorithms	Youtube comment	Over 95.4% in Logistic Regression

4.2. 8 SMS Spam

Despite the rapid development of Internet-based communication systems, SMS (Short Message Service) remains a vital tool for daily communication. Many companies feel that text messages are more effective than emails because of their higher open rates. Consumers only open one in every four emails they get, compared to 82% of SMS messages, according to research, which is read within the first five minutes. Spammers have, however, been drawn in due to the importance of SMS in the lives of mobile phone users. In recent years, the number of SMS spam has significantly expanded, posing new security risks like SMiShing.

In a study conducted by (Ghourabi, et al., 2020) To identify SMS spam, they developed a deep learning approach that is hybrid which targets mixed text messages that are authored in either Arabic or English. Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) are two deep learning techniques combined in this detection model. Two datasets were used to assess the efficacy of this method. The initial collection of data comes from the UCI Repository and consists of 5,574 English messages that have been classified as spam or valid (ham). The second set of data consisted of 2,730 Arabic texts that were classified as spam and not spam and were taken from various Saudi Arabian devices. They also tried a number of well-known machine-learning methods for comparative evaluation. he CNN-LSTM model surpasses the other methods, as shown by the experimental findings reported in this work, which show an outstanding precision rate of 98.37%.

Yerima & Bashar affirms that The number of SMS messages exchanged daily around the world has increased significantly during the last several years. Unfortunately, this surge in SMS usage has also increased the potential of SMS spam messages reaching mobile devices, which could result in fraud and the theft of user data. As a result, it is now essential to create message filtering systems to identify and remove SMS spam, and new machine learning techniques are continually being investigated for this purpose. They proposed a system in this study that uses a semi-supervised novelty detection strategy to identify SMS spam. By using a one-class Support Vector Machine (SVM) classifier, the system acts as an anomaly detector and learns only from regular SMS messages. Due to this innovative method, detection models can be used even in the lack of tagged SMS spam data for training purposes. They conducted trials using a reference data set made up of 747 spam SMS messages and 4,827 non-spam SMS messages in order to assess the performance of the suggested system. Based on TF-IDF (Term Frequency-Inverse Document Frequency) bag-of-words representations, frequency and binary, the results show that their method outperforms conventional supervised machine learning algorithms. The system's overall accuracy was 98%, with a 100% SMS spam identification rate and an exceptionally low 3% accidental detection rate. (Yerima & Bashar, 2022).

Table 8: SMS Spam

Paper	Task	Methods	Dataset used	Outcomes
(Ghourabi, et al, 2020)	Classification and Detection	Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM)	UCI spam repository for SMS	Hybrid CNN-LSTM has > 98% accuracy
(Yerima & Bashar, 2022)	Detection	Support Vector Machine (SVM)	Benchmark dataset	Over 98% accuracy

4.2.9 Proposed Spam-filtering Technique and Architecture

Our attention has thus far been on the numerous strategies developed or put out to combat different forms of spam on various social media sites. The creators and operators of social media platforms place a strong emphasis on anti-spam measures as spamming becomes a serious issue. The previous sub-sections described recent advancements in spam-fighting methods that have been published and are listed in the literature.

Likewise, we looked through the many developer blogs and other resources to see if there were any clues that could help us comprehend the fundamentals of the technology behind the algorithms for spam identification and filtering. Therefore, a Deep Learning technique is proposed which leverage on Neural network architectures for social spam detection to automatically recognise and categorise social media posts as spam or authentic. Here is a step-by-step description of the method:

- *Data Gathering:* Compile a sizable amount of labelled information that contains illustrations of social media posts that are both spam and not spam. The deep learning model will be trained and evaluated using this data.
- *Data cleaning and preprocessing:* Remove noise and unimportant data from the collected data. This could entail actions like eliminating stop words, changing the text's case, and removing punctuation. Tokenize the text and normalise the terms using methods like stemming or lemmatization.
- *Word Embeddings:* Use word embeddings, such as Word2Vec, GloVe, or FastText, to represent the pre-processed text. Word embeddings capture the semantic connections between words and give the neural network useful representations from which to learn.
- *Neural Network Architecture:* Create a neural network design that is appropriate for the goal of detecting social spam. Recurrent neural networks (RNNs), convolutional neural networks (CNNs), or a combination of the two may be used for this. While CNNs are excellent at identifying local patterns and features, RNNs excel at catching sequential dependencies.
- *Training, Testing/Validation:* Split the labelled data into training and testing/validation sets for the model. Utilise the training set to train the deep learning model. During training, use backpropagation and gradient descent algorithms to optimise the model's parameters. To avoid overfitting, take into account strategies like early halting and regularisation.
- *Model Evaluation:* Utilise the labelled testing/validation data to assess the trained model. Measure the model's performance in terms of accuracy, precision, recall, and F1 score to determine how well it can identify social spam.
- *Hyperparameter Tuning:* To improve the performance of the model, adjust the hyperparameters it uses. To find the ideal configuration, experiment with various network designs, activation functions, learning rates, batch sizes, and regularisation strategies.
- *Model Deployment:* Deploy the model to a production environment so that it may be used for actual spam detection when it has been trained and optimised. This could entail developing a separate application for spam detection or incorporating the model into an already-existing social media platform.
- *Continuous Improvement:* Track the effectiveness of the implemented model and gather user input. Improve and retrain the model frequently with fresh, labelled data to account for changing spamming tactics and enhance accuracy over time.

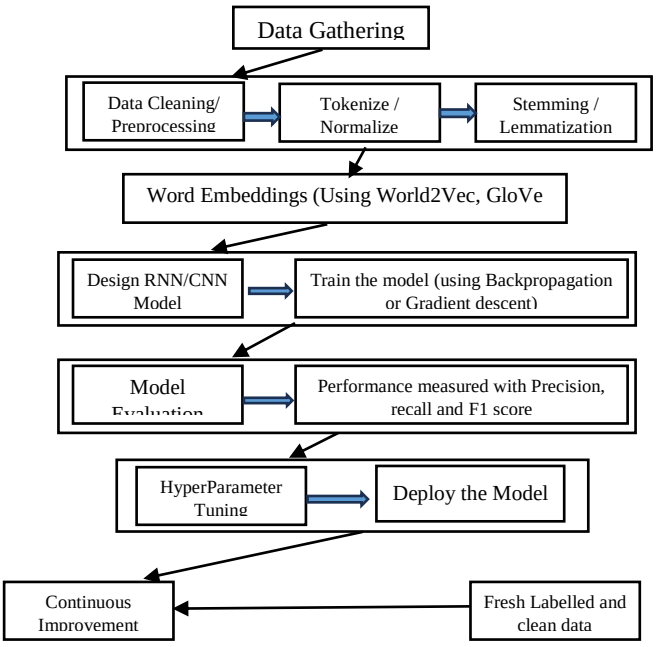


Figure 9: Proposed Spam Filtering Technique Architecture

This deep learning approach can be used to create a system for detecting social media spam that can accurately identify and categorise social media content as spam or legitimate, giving consumers a more secure and enjoyable experience on social platforms. The quantity and quality of the labelled data, the neural network architecture selected, the efficacy of the preprocessing techniques, and the precision of the hyperparameter tuning can all affect the expected outcomes when adopting the deep learning-based social spam identification method. It's vital to remember that the precise implementation, the calibre of the labelled data, and the difficulty of the spam detection task can all affect the final outcomes. The system must be monitored, evaluated, and updated frequently in order to be effective and continue to perform better over time.

5. CONCLUSION

Spamming has become a pervasive issue in our online existence, affecting various forms of media. Various methods for screening out spam have been investigated across platforms with various degrees of effectiveness. This study focuses especially on new strategies for social spam filtering. The paper opens with a summary of traditional approaches before delving into more recent advancements in social spam detection and mitigation across a variety of media outlets and applications, including emails, blogs, microblogs, SMS, and social networking. These methods fall into one of three categories: graph-based algorithms, probabilistic or deterministic, with each showing a lot of diversity within its own category. Comprehensive study demonstrates that, despite differences in their individual implementations, recent strategies primarily use probabilistic methodologies. This predilection can be related to social networks' distinctive features, where posts are often brief, personal, opinionated, filled with local references, and occasionally featuring cryptic and sarcastic content. In some cases, these posts may be incomprehensible even to humans, let alone automated systems. Consequently, it provides an intricate task to accurately detect and profile all the elements in social media information.

It is crucial to understand that the conflict amongst scammers and spam-fighters is a never-ending struggle. Spammers quickly reverse-engineer new detection algorithms as spam tactics change in order to get beyond filtering systems. Therefore, ongoing monitoring and the creation of better spam-fighting strategies are perpetual demands.

This study attempts to assemble a list of previously employed spam-combating strategies in social media, serving as a road map for subsequent efforts to combat this problem.

REFERENCES

Abinaya, R., E., B. N. & Naveen, P., 2020. Spam Detection On Social Media Platforms. *7th International Conference on Smart Structures and Systems (ICSSS),*, pp. 1-3.

Abu-Salih, B., Qudah, D. A., Al-Hassan, M. & Ghafari, S. M., 2022. *An Intelligent System for Multi-topic Social Spam Detection in Microblogging*, Jordan: arxiv.

al., J. e., 2021. *Detecting Nuisance Calls over Internet Telephony Using.* Limeric, Ireland: MDPI.

Al-Zoubi, A. M., Alqatawna, J. & Hassonah, M. A., 2019. Spam profiles detection on social networks using computational intelligence methods: The effect of the lingual context. *Journal of Information Science,* 47(1).

Antonov, E., Kontsewaya, Y. & Artamonov, A., 2021. Evaluating the Effectiveness of Machine Learning Methods for Spam Detection. *Procedia Computer Science,* Volume 190, pp. 479-486.

Balogun, A. K. et al., 2017. Spam Detection Approaches and Strategies: A Phenomenon. *International Journal of Applied Information Systems (IJAIS),* 12(9).

Binsaeed, K., 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(IJACSA),* 11(11).

Caha, T. & Kovaˇr´ık, M., 2022. Spam filter based on geographical location of the sender. *Electrical Engineering,* Volume 73, pp. 292-298.

Chaudhry, S., Dhawan, S. & Tanwar, R., 2020. *Spam Detection in Social Network Using*, Dehradun, India: Research Gate.

DigitalMaas, 2018. *How to Win the Fight Against Fake Reviews on Google.* [Online] Available at: https://www.digitalmaas.com/blog/win-fight-fake-reviews-google/ [Accessed 1 July 2023].

GCFGlobal, 2023. *Using Search Engine.* [Online] Available at: https://edu.gcfglobal.org/en/internetbasics/using-search-engines/1/# [Accessed 2 July 2023].

Ghourabi, A., Mahmood, M. A. & Alzubi, Q. M., 2020. A Hybrid CNN-LSTM Model for SMS Spam Detection in Arabic and English Messages. *Future Internet - Journal of Cybersecurity,* 12(9).

Gong, Q. et al., 2018. *DeepScan: Exploiting Deep Learning for Malicious Account Detection in Location-Based Social Networks,* s.l.: IEEE Communications Magazine.

Hayoung, O., 2021. A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model. Volume 9, pp. 144121 - 144128.

He, X., Gong, Q., Chen, Y. & Fu, X., 2021. *DatingSec: Detecting Malicious Accounts in Dating Apps Using a Content-Based Attention Network,* s.l.: Institute of Electrical and Electronics Engineers.

Hossain, F., Uddin, M. N. & Halder, R. K., 2021. *Analysis of Optimized Machine Learning and Deep Learning Techniques for Spam Detection,* Toronto, ON Canada: IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS).

Hussain, N. et al., 2021. Detecting Spam Product Reviews in

Roman Urdu Script. *The Computer Journal (IF 1.762),* 64(3), p. 432–450.

ICTEA, 2023. *What is the unsolicited email (spam)?.* [Online] Available at: https://www.ictea.com/cs/index.php?rp=%2Fknowledgebase%2F2063%2FiQue-el-correo-no-deseado-spam.html&language=english [Accessed 2 July 2023].

Jain, A. K., Sahoo, S. R. & Kaubiyal, J., 2021. Complex and Intelligent System. *Online social networks security and privacy: comprehensive review and analysis,* pp. 2157 - 2177.

Javed, I. et al., 2021. *Detecting Nuisance Calls over Internet Telephony Using.* Electronics 2021, 10, 353. ed. Basel, Switzerland: MDPI.

Kabakus, A. T. & Kara, R., 2019. "TwitterSpamDetector": A Spam Detection Framework for Twitter. *International Journal of Knowledge and Systems Science (IJKSS) ,* 10(3), p. 14.

Kardaş, B., Bayar, İ. E., Özyer, T. & Alhajj, R., 2021. Detecting spam tweets using machine learning and effective preprocessing. *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* Volume 393-398.

Koggalahewa, D., Xu, Y. & Foo, E., 2021. Unsupervised Spammer Detection in Social Networks based on User Information Interests. *Research Square.*

KRTV, 2022. *Spammers are using bogus video links in comments.* [Online] Available at: https://www.krtv.com/news/spammers-are-using-bogus-video-links-in-comments [Accessed 7 July 2023].

Leslie, P., 2022. *Zero Bounce.* [Online] Available at: https://www.zerobounce.net/blog/email-resources/be-a-better-marketer/the-origin-of-the-word-spam-email-spam#:~:text=Some%20say%20it's%20an%20acronym,the%20television%20screen%20in%201970. [Accessed 1 July 2023].

Li, M., Wu, B. & Wang, Y., 2019. *Comment Spam Detection via Effective Features Combination,* Shanghai, China: IEEE.

N. Hussain, H. et al., 2020. Spam Review Detection Using the Linguistic and Spammer Behavioral Methods. *IEEE Explore,* Volume 8, pp. 53801-53816.

NGUYEN-KHOA-MAN, N., 2022. *Detecting spam on a blog platform: a machine-learning approach,* s.l.: Webedia I/O.

Orika, J. T., 2022. *Foundation Lab.* [Online] Available at: https://foundationinc.co/lab/social-media-spam [Accessed 06 07 2023].

Reddit, 2022. *Reddit.* [Online] Available at: https://www.reddit.com/r/mildlyinfuriating/comments/xdw0rp/twitter_spam_message_requests/ [Accessed 8 July 2023].

Riserbato, R., 2019. *The What, Why, & How of Social Bookmarking.* [Online] Available at: https://blog.hubspot.com/marketing/social-bookmarking

Samsudin, N. M. et al., 2019. Youtube spam detection framework using naïve bayes and logistic regression. *Indonesian Journal of Electrical Engineering and Computer Science ,* 14(3), pp. 1508 - 1517.

Sanjeev, R., Verma, A. K. & Bhatia, T., 2021. A review on social spam detection: Challenges, open issues, and future directions. *Expert Systems with Applications,* Volume 186.

Shahariar, G. M. et al., 2022. Spam Review Detection Using Deep Learning. *AirXIV,* Volume 1.

Shahzad, A., Mahdin, H. & Nawi, N. M., 2020. An Improved Framework for Content-based Spamdexing Detection. *(IJACSA) International Journal of Advanced Computer Science and Applications,,* 11(1), p. 51.

Shen, H., Liu, X. & Zhang, X., 2022. A Detection Method for Social Network Images with Spam, Based on Deep Neural Network and Frequency Domain Pre-Processing. *Cyber Security and Critical Infrastructures,* 11(7).

SproutSocial, 2023. *Sprout Social.* [Online] Available at: https://sproutsocial.com/glossary/microblog/ [Accessed 5 July 2023].

Statista, 2022. *Global spam volume as percentage of total e-mail traffic from 2011 to 2022.* [Online] Available at: https://www.statista.com/statistics/420400/spam-email-traffic-share-annual/ [Accessed 8 July 2023].

Sultana, T., Sapnaz, K. A., Sana, F. & Najath, M. J., 2022. Email based Spam Detection. *International Journal of Engineering Research & Technology (IJERT),* p. Vol. 9 Issue 06.

Sumathi, M. & Raja , S. P., 2023. *Machine learning algorithms-based spam detection in social networks,* Thanjavur, India: Research Square.

Techopedia, 2023. *50+ Phishing Statistics You Need to Know – Where, Who & What is Targeted.* [Online] Available at: https://www.techopedia.com/phishing-statistics [Accessed 8 July 2023].

Tolentino, J., 2015. *5 types of social spam (and how to prevent them).* [Online] Available at: https://thenextweb.com/news/5-types-of-social-spam-and-how-to-prevent-them [Accessed 25 June 2023].

Weisen, P., 2022. *Semantic Graph Neural Network: A Conversion from Spam Email Classification to Graph Classification.* Beijing, China: Hindawi.

Wikipedia, 2023. *Wikipedia.* [Online] Available at: https://en.wikipedia.org/wiki/Spamming [Accessed 1 July 2023].

Yerima, S. Y. & Bashar, A., 2022. *Semi-supervised novelty*

detection with one class SVM for SMS spam detection, Sofia, Bulgaria: International Conference on Systems, Signals and Image Processing (IWSSIP).

Zavrak, S. & Yilmaz, S., 2022. Email Spam Detection Using Hierarchical Attention Hybrid Deep Learning Method. *Machine Learning (cs.LG); Neural and Evolutionary Computing (cs.NE),* Volume 2.