SPAM DETECTION FOR EMAIL FILTERS USING ARTIFICIAL INTELLIGENCE TECHNIQUES

Afaf Abdu Yahya Al-Nowidi [1]
Abdulaziz Ahmed Thawaba [1*]
Maqbol Ahmed [2]

Received: 20/09/2025
Revised: 29/I0/2025
Accepted: 20/12/2025

[1] Department of Computer Science, Faculty of Information Technology and Computer Science, University of Saba Region, Marib, Yemen.
[2] Department of Computer Information Systems, Faculty of Information Technology and Computer Science,University of Saba Region, Marib, Yemen.
*Corresponding Author's Email: azizth@usr.a

Spam Detection for Email Filters Using Artificial Intelligence Techniques

Afaf Abdu Yahya Al-Nowidi
Department of Computer Science, Faculty of Information Technology and Computer Science, University of Saba Region, Marib, Yemen.
Email: alnowidiafaf@gmail.com

Abdulaziz Ahmed Thawaba
Department of Computer Science, Faculty of Information Technology and Computer Science,University of Saba Region, Marib, Yemen.
Email: azizth@usr.a

Maqbol Ahmed
Department of Computer Information Systems, Faculty of Information Technology and Computer Science,University of Saba Region, Marib, Yemen.
Email: Maqbol3@usr.ac

Abstract— Spam emails represent a significant threat to digital communications, compromising user privacy and security. Technological advancements have made traditional filtering methods, such as blacklists, conventional machine learning classifiers, and rule-based methods, incapable of adapting to the sophisticated techniques of spammers. This research focuses on providing a systematic analysis of techniques and models used in spam detection. It also examines how these models, such as artificial neural networks (ANNs), extract text features using TF-IDF and classify them to capture complex and nonlinear patterns in email data. In this research, it was found that some of the models proposed by the researchers for spam detection outperform traditional classifiers. The study demonstrated that the hybrid model proposed by the researchers, combining natural language processing and artificial neural networks, exhibits superior performance. When TF-IDF-based feature extraction is combined with an artificial neural network, this approach achieves higher accuracy and more robust semantic representation, enabling it to detect complex linguistic patterns common in sophisticated spam messages. Although the hybrid model requires higher computational costs and greater sensitivity to linguistic variation and class imbalances, its superiority over all other techniques justifies this trade-off. In contrast, models such as artificial neural networks alone, simple Bayesian algorithms, and support vector machines demonstrate acceptable performance but remain limited by either weak semantic understanding or overly robust modelling assumptions.

Keywords— *Email, Spam Detection, Email Filters, Artificial Neural Networks (ANN), Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP).*

I. INTRODUCTION

Email was the first publicly available means of electronic communication and has evolved significantly since its inception. Today, it is an integral part of various modern electronic services. (Rashid & Bhat, 2023). Email allows users to easily perform multiple tasks, all within a single environment. Furthermore, the rapid development of smart devices, such as phones and watches, has made email available anytime, anywhere, making it an essential tool for both business and personal communication. The proliferation of unsolicited emails (spam) poses a significant challenge, especially as our daily activities become increasingly intertwined with technology and business. These unwanted messages range from malicious phishing attempts to harmless advertisements, which can negatively impact productivity (Kourentzes et al., 2022). Spam is a collection of messages sent over the internet that may be unwanted or inappropriate and often clog users' inboxes. Spam clutters a user's inbox and makes it difficult to search official correspondence for important information, leading to frustration and a reduced user experience (Thakur et al., 2023). Early detection using effective filtering mechanisms to separate spam from legitimate emails is crucial because it allows users to focus on what matters to them. Therefore, it has become essential to develop effective email filters capable of detecting and separating spam from legitimate messages (Ali et al., 2021). Researchers have improved several filters that use artificial intelligence techniques to analyze incoming messages and identify spam patterns. By using these techniques, email systems can enhance their ability to distinguish between desirable and undesirable content. Furthermore, AI techniques improve the performance of filters over time, allowing them to adapt to evolving anti-spam methods (M. S. Abbasi, 2023).

Research has shown that traditional spam and phishing detection techniques face significant challenges due to the rapid evolution of malicious methods. Despite the application of modern technologies such as Natural Language Processing (NLP) and Artificial Neural Networks (ANNs), these techniques still suffer from errors in correctly classifying messages, resulting in frequent failures to detect malicious content effectively (Kourentzes et al., 2022). Numerous studies have emphasized the importance of using up-to-date datasets

that reflect current threats and not relying on outdated data, as this may reduce the effectiveness of detection models and weaken the ability to monitor changing patterns in phishing techniques (Wang & Zhao, 2021). Furthermore, other research has confirmed that attack methods and tactics are constantly evolving, which necessitates the development of new detection technologies that are flexible and intelligent enough to adapt to these changes (Al-Garadi et al., 2016). This study will attempt to highlight a number of spam and phishing detection techniques that utilize modern artificial intelligence technologies, while discussing aspects of excellence and shortcomings in these techniques according to a systematic review of published research in this field.

II. RESEARCH METHODOLOGY

As a methodological framework for this study, a comprehensive systematic literature review was conducted to establish the theoretical and practical foundation. This research reviewed more than 80 articles from peer-reviewed journals and research papers published at international conferences to gain a comprehensive understanding of current trends and challenges in the field of spam detection (Özbay, 2023).

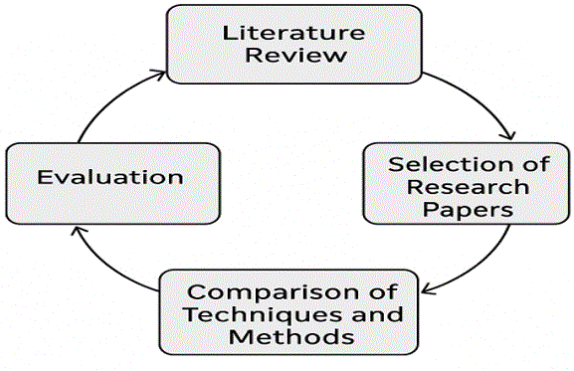

Figure 1: Research Methodology Processes

Figure 1 illustrates the research framework by selecting a subset of highly relevant research, published between 2020 and 2025, from a set of 87 literature based on their methodological rigor, direct relevance to AI-based spam filtering, and contributions to the field. As shown in Table 1, these studies were then analyzed and compared, focusing on the techniques, models, and evaluation strategies used in each. The analysis then encompassed a range of methodologies, ranging from classical machine learning algorithms (Naive Bayes and Support Vector Machines) to deep learning techniques and hybrid models. The final phase of the methodology in this research is to conduct a systematic comparison to uncover valuable insights into the strengths and weaknesses of each technology or hybrid technology, highlight gaps, and discover performance patterns across different datasets (Özbay, 2023).

Table 1 Selected References

Source	Definition	Strategy	Findings	Limitations
Abbasi (2023)	PhD thesis on ransomware detection using ML	Behaviour-based ML classification	Effective in detecting evolving threats	Focused on ransomware, not spam
Ali et al. (2021)	ML-based cyberbullying detection	Hyperparameter optimisation + oversampling	Improved ML performance in imbalanced datasets	Domain-specific (not spam)
AmeerSaleem (2025)	Evaluation of classification metrics	Review of metrics (Precision, Recall, F1)	Clarified the role of evaluation metrics	Secondary review; not a new spam method
Anonymous (2020)	Spam detection using NLP & DNN	ANN + NLP integration	Reduced false positives and improved accuracy	Computationally intensive
Hassani et al. (2020)	Text mining in big data analytics	Big data text mining	Potential applications in spam detection	Not directly applied to spam
Kourentzes et al. (2022)	Hybrid models (DL + NLP) for spam detection	Hybrid deep learning & NLP	Outperformed traditional methods	High computational cost
Liu & Huang (2020)	Ensemble learning for spam detection	Ensemble of multiple algorithms	Higher accuracy & robustness vs. single models	Requires more training time/resources
Liu et al. (2023)	Unsupervised DL for time series	Unsupervised deep learning	Detected complex anomalies	Not specific to spam detection
Özbay (2023)	DL for image classification	Active DL + optimisation	Improved classification accuracy	Domain-specific, not spam
Rashid & Bhat (2023)	DL for influencer detection in OSNs	Systematic review	DL advances are transferable to spam detection	Domain-specific
Shehnepoor et al. (2023)	Fraud detection review	Analysis of ML/DL fraud detection	Identified gaps in fraud analysis	Not applied to spam
Slodkowski et al. (2023)	Review of recommender systems	Literature survey	Hybrid methods improved recommendations	Education-focused, not spam
Spark (2023)	NLP + Multinomial Naïve Bayes	Spark-based NLP + MNB	Efficient for big datasets	Weak against obfuscated spam
Talpur & O'Sullivan (2020)	Cyberbullying detection in Twitter	Feature engineering for text classification	Improved performance with imbalance handling	Domain: cyberbullying, not spam
Thakur et al. (2023)	Review of DL for phishing email detection	Systematic DL-based review	Showed DL advantages over ML	Phishing-specific
Wang & Zhao (2021)	Review of spam detection methods	Survey of ML + traditional methods	Highlighted spam evolution & dataset issues	No novel experiments

III. THEORETICAL BACKGROUND

3.1 Email & SPAM

Email is the most widely used and important form of digital communication, facilitating the exchange of information across computer networks and smart devices. The widespread use of email is due to many factors, including its cost-effectiveness, speed, reliability, and ease of use, as well as the availability of email services offered by platforms such as Hotmail, Yahoo, and Gmail, as well as by Internet Service Providers (ISPs) (Thakur et al., 2023). SPAM, also known as junk email, is unsolicited email sent in large quantities without the recipient's consent. SPAM is unsolicited mass emails posted online for advertising, phishing, or malicious purposes (Ali et al. 2021, 2021). As digital communication proliferated, spam evolved beyond simple promotional content to include sophisticated phishing schemes, malware distribution, and fraudulent activities. The impact of spam extends beyond mere nuisance; it imposes considerable burdens on users and network infrastructure. Spam consumes bandwidth, occupies storage space, and frequently carries harmful payloads that compromise recipient devices. Moreover, infected systems can be covertly transformed into spam-sending entities, further exacerbating the spread without the user's awareness (A. Abbasi & Chen, 2008; Hassani et al., 2020; Talpur & O'Sullivan, 2020).

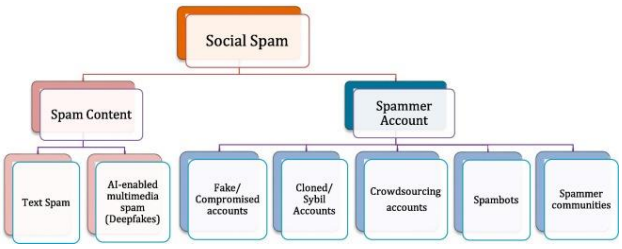

Figure 2: Spam Emails (Shehnepoor et al., 2023)

As shown in Figure 2, spam, known as junk mail, is defined in different ways; these definitions explain the difference between spam and legitimate mail (also referred to as ham, non-spam, or genuine mail). Among the most common definitions, the shortest one defines spam as "unsolicited bulk email." (Ali et al., 2021). As the internet grew, so did spam, evolving from simple promotional emails to phishing scams, malware distribution, and fraud.

3.2 Spammers

Spammers employ a variety of techniques to conceal their identities when disseminating spam emails. Interestingly, while they meticulously obscure their traces during the distribution phase, they often fail to hide their identities during the harvesting of email addresses, which typically involves extracting contact information from publicly available sources, such as websites and academic papers. One of the most telling aspects of spam activity is the way in which it evolves in response to defensive measures. Spammers rarely remain static; instead, they constantly modify their strategies to bypass newly introduced filters and detection methods (Özbay, 2023). This adaptive behavior means that even the most advanced filtering solutions often have a limited period of effectiveness, which in turn places pressure on researchers and practitioners to update and refine detection models continuously. Earlier studies, such as those by Pu and Webb, highlight how the structural patterns that spammers rely upon eventually become less effective once they are confronted with sophisticated countermeasures (Shehnepoor et al., 2023). In other words, there is a persistent cycle of innovation and counter-innovation, where every advance in filtering technology prompts spammers to devise fresh evasion tactics. This ongoing contest emphasizes the importance of building detection frameworks that are not only accurate but also flexible and capable of evolving alongside the ever-changing landscape of spam generation (Hassani et al., 2020, 2020; Talpur & O'Sullivan, 2020).

3.3 Spammers' Tricks

To mount large-scale spam campaigns, adversaries first need extensive address lists. In practice, these lists are assembled in two main ways: automated harvesting and commercial acquisition. Harvesting tools crawl public web resources, including websites, forums, comment sections, and online directories, to extract exposed email addresses at scale (Ali et al., 2021). Meanwhile, other operators expand their reach by purchasing or leasing pre-compiled lists from third-party brokers within the spam ecosystem. Armed with large recipient pools, spammers then employ a range of evasive measures designed to circumvent both content-based and network-level defenses. These tactics include textual obfuscation, header forgery, botnet distribution, IP rotation through offshore or compromised hosts, and exploitation of misconfigured mail relays or third-party posting services. The objective is consistent: maximize delivery while minimizing detection and attribution (Al-Garadi et al., 2019). For clarity, Table 2 summarizes the most prevalent evasion techniques observed in the literature and in operational analyses, together with brief descriptions of how each method subverts common filtering mechanisms.

Table 2 Common Tricks Used by Spammers to Evade Detection (Al-Garadi et al., 2016)

Trick	Description
Zombies or Botnets	Leveraging compromised computers to distribute vast quantities of spam, malware, or viruses across the internet.
Bayesian Sneaking & Poisoning	Incorporating atypical words or structures in spam content to confuse Bayesian filters and prevent accurate classification.
IP Address Manipulation	Utilising or acquiring IP addresses with reputable or neutral histories to bypass reputation-based filters.
Offshore ISPs	Routing spam through Internet Service Providers in jurisdictions with lax security enforcement.
Open Proxies/Open Relays	Exploiting misconfigured servers to relay spam, masking the true origin of the messages.
Third-party Mailback Software	Abusing poorly secured mailing functionalities on legitimate websites to send spam.
Falsified Header Information	Inserting deceptive or fake metadata into email headers to obscure the message's source.
Obfuscation	Using misleading characters or HTML tags to fragment words and evade keyword-based filters.
Vertical Slicing	Structuring spam content vertically to disrupt standard parsing techniques.
HTML Tricks	Employing HTML formatting to disguise malicious content and avoid signature-based detection.

These techniques incorporate the adaptive and flexible nature of spammers, who constantly improve their methods in response to advances in filtering technologies. This competition highlights the urgent need for dynamic and intelligent spam detection models capable of recognizing not only static indicators of SPAM but also sophisticated evasion strategies (Slodkowski et al., 2023).

IV. EVOLUTION OF SPAM TECHNIQUES

Spammers have adapted their methods to circumvent detection systems, using techniques such as obfuscation, where important keywords are deliberately distorted (e.g., "free" spelled as "fr3e"), to evade keyword filters. Additionally, social engineering has become widespread, where spam messages include malicious links or attachments disguised as legitimate content. (Ali et al., 2021). Early approaches to combating spam were grounded in static filtering methods, most notably blacklists and whitelists, which categorized emails according to predefined sender addresses, domains, or reputational scores. While straightforward to implement and initially effective, such techniques quickly revealed critical limitations. The fast-paced evolution of spam campaigns, coupled with the constant compromise of new devices and servers through malware, meant that spammers could rapidly shift to fresh infrastructures that were not yet recorded on static lists. Therefore, blacklists and whitelists become outdated as soon as they are updated, making these defenses powerless against the sophisticated and adaptive strategies used by spammers. This has highlighted the need to move from static mechanisms to dynamic, flexible, and intelligent adaptive filtering solutions (Slodkowski et al., 2023).

Dynamic filtering strategies rely on analyzing email content, structure, and language patterns and use machine learning algorithms and statistical models to identify anomalies that indicate spam (Al-Garadi et al., 2016; Özbay, 2023). These techniques provide greater flexibility by analyzing data and identifying subtle signals that static lists cannot recognize. However, the results demonstrate that they are not without flaws. False positives can disrupt legitimate communications, while false negatives can allow malicious messages to leak. The effectiveness of these models also depends heavily on the availability of large, up-to-date datasets that reflect the changing tactics of spammers. To

address these limitations, recent studies have increasingly emphasized the integration of ML with NLP to build more sophisticated and accurate models. By combining advanced text representation techniques with powerful classifiers, these models improve accuracy and generalization across various spam scenarios. These models hold great promise in overcoming the weaknesses of previous techniques and provide a path toward more flexible and intelligent spam filtering solutions (Kourentzes et al., 2022).

4.1 Existing Solutions for Spam

Previous studies have shown several ways to mitigate the impact of spam and enhance the security and reliability of email communication. Anti-spam methods typically rely on filters designed to protect the user by intercepting unwanted messages before they arrive and preventing phishing attacks or malware (Ali et al., 2021). Table 3 summarizes selected studies that have contributed to the development of anti-spam techniques. It highlights the methodology proposed in each study, provides valuable insights into the ongoing development of spam detection techniques, and reveals some of the limitations inherent in their methods.

Table 3 Existing Anti-Spam Methodologies

Author	Description
Sankar et al.	In this study, a method for detecting disguised spam messages is presented by completing synonym relationships and linking keywords. This approach predicts the spam category by analysing the message content, bypassing reliance on static keyword lists.
Bhowmick et al.	The study focused on a comprehensive systematic review of several ML-based spam filters and their different types. The study evaluated existing techniques, assessed their effectiveness, and documented developments in the field of spam filtering.
Mirza et al.	Here, a spam classification system is developed using a naive Bayesian classifier with a focus on distinguishing between spam and non-spam classes, based on content analysis.
McGetrick et al.	Here, IBM Watson's Tone Analyser API was used to extract personality insights and language tone scores from emails. These features were then used in machine learning models to classify emails as spam or real messages.
Bassiouni et al.	This study focused on investigating spam classification using 10 different classifiers on a benchmark dataset, which helps in comparative performance evaluations to identify suitable algorithms.

All of these diverse strategies and techniques together represent a broad spectrum of efforts to mitigate the spam problem, ranging from linguistic analysis and statistical classification to more sophisticated ML frameworks. Despite the significant development of these approaches, they still face challenges in adapting quickly enough to the changing behavior of spammers (Wang & Zhao, 2021). They rely heavily on manual or highly specific feature engineering and may suffer from high rates of false positives and false negatives, which undermines their reliability in practical applications. This scenario highlights the urgent need to continue research into accurate, dynamic, and adaptive detection models. Future methodologies and models must be able to evolve in parallel with adversarial behavior, narrowing the gap between the emergence of new spam techniques and the development of filtering technologies. Advances in Natural Language Processing (NLP) and Machine Learning (ML) present particularly promising prospects, as they enable richer contextual understanding and enhanced resistance to obfuscation (Wang & Zhao, 2021). With these advancements, next-generation spam detection frameworks can enhance their power and accuracy, providing more effective protection against the ever-changing threat of unwanted and malicious communications.

4.2 Spam Classification

Data classification is a critical process in spam detection systems, enabling the accurate classification of emails into spam or legitimate (important) categories, as illustrated in Figure 3. A variety of machine learning algorithms have been used for this purpose, including artificial neural networks (ANNs), Naive Bayes, and support vector machines (SVMs). These techniques are designed to examine the content of email messages, extract relevant features, and identify underlying patterns that indicate spam (Kourentzes et al., 2022). The effectiveness of these classifications is typically evaluated using performance metrics such as precision, recall, and F1 score, which collectively measure the

balance between accurately identifying spam and reducing false alarms. Despite significant progress, spam detection remains a complex challenge within the scope of traditional classification challenges. This complexity is based on the "indistinguishability theorem," which assumes that no single classification method outperforms any other across all types of data. Therefore, the ongoing search for alternative models and improvement strategies is crucial. Several recent studies have explored new approaches and techniques for spam classification. For example, researchers have proposed a logistic regression model trained using metaheuristic algorithms, along with a fine-tuned XGBoost model using similar optimization strategies (Wang & Zhao, 2021).

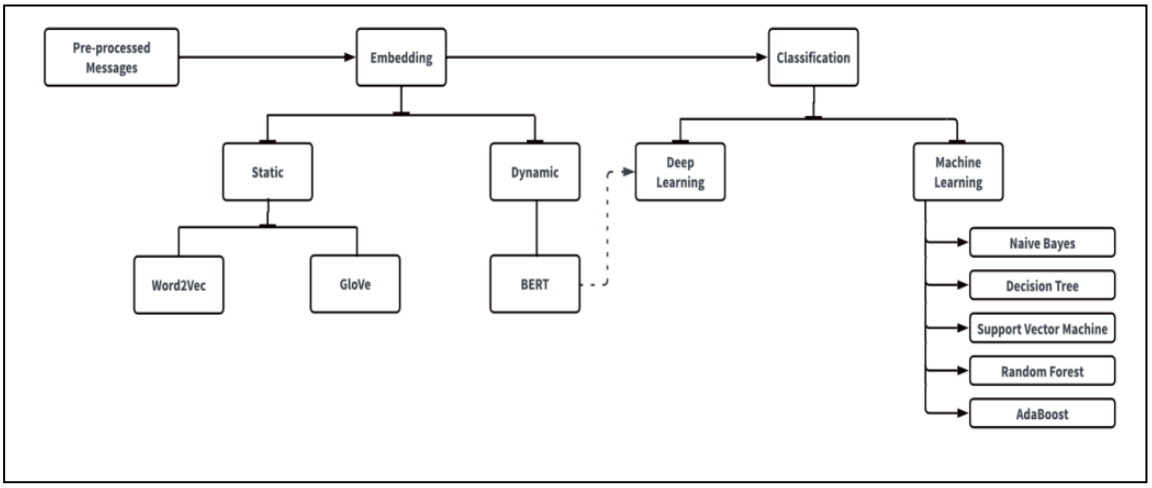

Figure 3: Flow diagram of spam classification

4.3 Classification Pipeline

As illustrated in Figure 3, a standard spam classification pipeline begins with preprocessing raw email data. This stage involves cleaning the textual content by stemming words to their roots and removing punctuation. The cleaned words are then transformed into vector embeddings, preserving their semantic meaning and enabling machine learning algorithms to process them efficiently. These embeddings serve as the main input for further analysis and classification activities.

Text Classification: Abbasi and Chen (2008) demonstrated that text classifiers rely on analyzing email content using several techniques, such as bag-of-words and TF-IDF representations. These techniques transform text data into structured digital vectors, facilitating the detection of suspicious content patterns.

Metadata and Behavioral Classification: Al-Jaradi et al. (2016) highlighted the significance of integrating metadata features, including sender reputation, frequency of sending, and atypical attachment behaviors, as these factors frequently indicate spam activity that might not be evident through content analysis alone.

Handling Dynamic Challenges: The dynamic nature of spam remains a major obstacle in email detection. As recent studies on spam detection using improved NLP and deep neural networks (DNN) techniques have demonstrated, spammers are constantly adapting and changing their tactics by altering language structures and concealing malicious content. This necessitates continuous updates to classification models to maintain their effectiveness.

V. TRADITIONAL APPROACHES

Traditional approaches have focused on basic spam detection approaches, particularly blacklists and content-based filtering. Although these approaches are effective in stable environments, they often falter when faced with constantly changing spam messages. Although blacklists can block known malicious senders, they have been largely ineffective against newly generated or fraudulent email addresses (X. Liu & Huang, 2020). Blacklists, which maintain records of recognized malicious email addresses or domains, are effective at blocking previously identified sources of spam; however, they have been largely ineffective against newly generated or fraudulent addresses. Content-based filtering examined the textual content of emails and recognized messages that include suspicious words, phrases, or links. Although effective for detecting spam regardless of the sender, this filtering method often produced false positives when legitimate emails shared similar keywords (Shehnepoor et al., 2023). The rule-based approach applied predefined criteria (e.g., a high number of links, unusual message length) to categorize emails. Nevertheless, these rules were rigid and could not adjust to the rapidly evolving strategies of spammers. While these approaches offered preliminary solutions for spam mitigation, their static nature made them insufficient to counter the increasingly sophisticated and adaptable strategies employed in contemporary spam campaigns. This limitation ultimately spurred a shift toward machine

learning and artificial intelligence techniques (Y. Liu et al., 2023; Wang & Zhao, 2021).

VI.SPAM DETECTION USING AI TECHNIQUES

Spam detection systems and software have evolved significantly and continuously over the years, moving from simple, rule-based methods to sophisticated machine learning-based solutions. Initial detection mechanisms relied heavily on static methods, such as keyword matching and pre-built blacklists. While initially effective, these systems quickly revealed their shortcomings when confronted with increasingly sophisticated and adaptive spam techniques. Contemporary spam detection frameworks employ dynamic learning models capable of recognizing complex patterns in email content. This flexibility enhances the ability to accurately distinguish legitimate emails from spam. Using statistical and machine learning algorithms, these systems outperform the static properties of traditional filters (Dada et al., 2019).

Taken together, these diverse strategies represent a broad spectrum of initiatives aimed at addressing the spam problem, ranging from basic linguistic analysis and statistical classification to more sophisticated ML frameworks. Many approaches struggle to keep up with the ever-evolving methods of spammers, relying heavily on manual or domain-specific feature engineering or encountering high rates of false positives and false negatives that undermine their effectiveness in practical applications (Wang & Zhao, 2021). This scenario highlights the urgent need to continue research into accurate, adaptive, and self-evolving detection models. Future systems must be able to keep pace with adversarial behavior, bridging the gap between the emergence of new spam methods and improved filtering techniques. Advances in natural language processing and deep learning offer highly encouraging opportunities, facilitating deeper contextual understanding and enhancing resistance to obfuscation. By leveraging these developments, next-generation spam detection frameworks can achieve greater robustness and accuracy, ensuring more effective protection against the ever-evolving threat of unwanted and malicious communications (A. Abbasi & Chen, 2008; Al-Garadi et al., 2016; Ali et al., 2021; Özbay, 2023). Modern Systems: Advances in artificial intelligence and machine learning have revolutionized spam detection practices. Abbasid and Chen (2008) demonstrated how models such as naive Bayes and support vector machines (SVMs) have significantly improved detection accuracy by learning from large datasets. Figure 4 illustrates how natural language processing can be used to detect spam in emails.

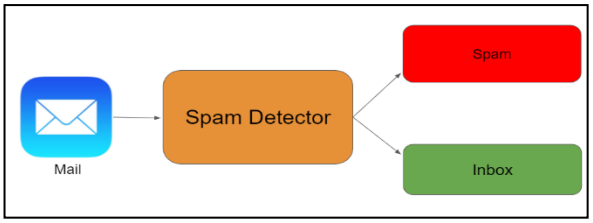

Figure 4: Using Natural Language Processing for Spam Detection in Emails (Kourentzes et al., 2022)

Accordingly, the incorporation of Natural Language Processing (NLP) enables systems to understand subtle contextual details within email content and thus identify subtle spam indicators. (A. Abbasi & Chen, 2008; Spark, 2023).

Hybrid and Ensemble Models: We have shown in recent work, such as “Spam Detection Based on Enhanced Natural Language Processing (NLP) and Deep Neural Networks (DNNs)," the effectiveness of hybrid systems that integrate machine learning with deep learning architectures. These frameworks often combine NLP-based text analysis with insights into network behavior, achieving superior detection rates by exploiting the strengths of diverse methodologies. (Anonymous, 2020)

Natural Language Processing (NLP) has become a key component in improving the efficiency of spam detection through various methodologies, with numerous studies highlighting its significant impact. Feature Extraction and Representation Approaches, such as TF-IDF and word embedding, enable the identification of important terms and semantic links in email texts, enabling models to more accurately distinguish spam from legitimate content. In Handling Multilingual Content: Various tools, including mBERT and LASER, provide language-independent representations, ensuring that detection systems remain effective across different linguistic environments. Dimensionality Reduction: Techniques such as Principal Component Analysis (PCA) optimize feature spaces, increasing computational efficiency and boosting the performance of machine learning classifiers. Feature Extraction and Representation: Approaches such as Term Frequency-Inverse Document Frequency (TF-IDF) help identify important terms in email content. By assessing the importance of a word in relation to its frequency, models can extract key features, facilitating the effective discrimination between spam and legitimate messages (A. Abbasi & Chen, 2008). Handling Multilingual Content: Advanced tools such as Multilingual BERT (mBERT) and LASER have been used to develop language-independent text representations. These techniques ensure that spam detection systems remain consistent across diverse linguistic contexts, accounting for the global nature of email communications. (Al-Garadi et al., 2016). Dimensionality Reduction: This method applies techniques such as principal component analysis (PCA)

along with TF and TF-IDF to reduce the dimensionality of the extracted features. This not only improves computational efficiency but also enhances the performance of subsequent classification models, such as artificial neural networks (ANNs) (A. Abbasi & Chen, 2008; Ali et al., 2021). In general, natural language processing supports three essential aspects of modern spam detection: structuring raw text data into parsable formats, enabling cross-language adaptation, and simplifying feature sets to improve model results.

6.1 Neural Networks (ANN)

Neural network methodologies have played a pivotal role in developing spam detection capabilities. Figure 5 illustrates how neural network techniques work. As evidenced by numerous recent studies, neural network methodologies include:

Deep Neural Networks (DNNs): This method is used to process high-dimensional features derived via TF-IDF, where DNNs skillfully capture complex spam patterns, resulting in increased classification accuracy. (A. Abbasi & Chen, 2008).

Recurrent Neural Networks (RNNs) and LSTM: Taking advantage of the sequential nature of email texts, recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, have been used to analyze multilingual data streams. This improves the system's capability to detect spam more accurately within various linguistic frameworks (Al-Garadi et al., 2016).

Integrated Models Utilizing Artificial Neural Networks (ANNs): Additional research has integrated ANNs to concurrently analyze both textual and static characteristics, such as metrics of sender reputation, resulting in the creation of effective scoring-based spam filtering systems. This holistic strategy has greatly enhanced detection efficacy. These neural network methodologies illustrate the vital connection between artificial intelligence and spam detection, emphasizing the proficiency of ANNs, recurrent neural networks (RNNs), and LSTM memory architectures in identifying intricate linguistic and behavioral patterns in spam. The synergistic application of these models alongside natural language processing (NLP) tools bolsters their function in developing spam-resistant and adaptive systems (X. Liu & Huang, 2020).

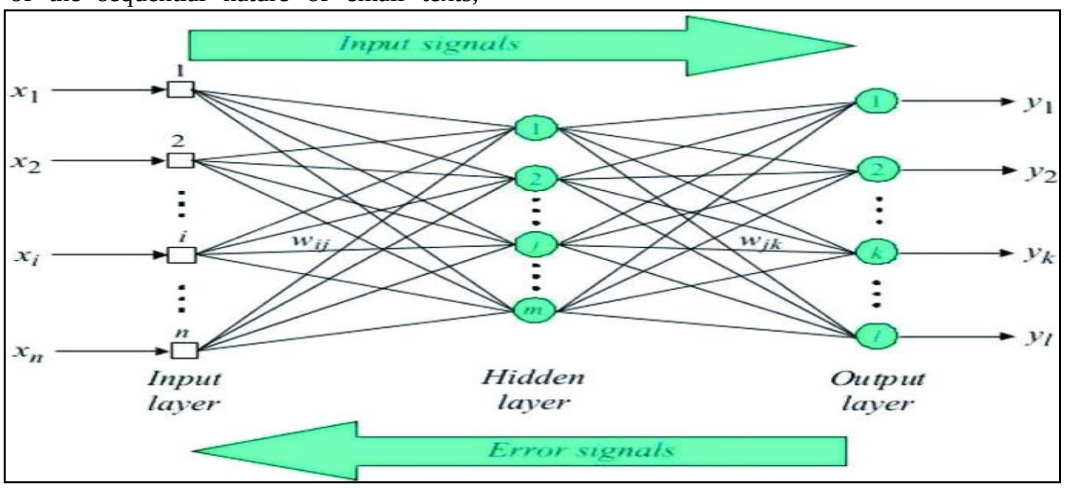

Figure 5: Neural network techniques

6.2 Feature Subset Selection

Selecting key features is a crucial step in developing effective and accurate spam detection models. By identifying and retaining only the most valuable features, computational costs can be reduced and the overall effectiveness of classification systems enhanced. Efficient selection of feature subsets also mitigates the risks associated with overfitting or underfitting, ultimately improving the generalizability of spam detection models (Slodkowski et al., 2023).

Challenges in Feature Subset Selection

Despite its effectiveness and importance, the selection of a subset of features in spam detection is fraught with many challenges that can negatively impact system performance (Ali et al., 2021):

High Dimensionality: Email data often includes a wide range of features, including metadata, textual content, and sender behavior. These high dimensions place a significant burden on traditional algorithms, complicating both processing and analysis.

Irrelevant or Redundant Features: Many of the extracted features may have little effect on distinguishing spam

from legitimate emails. In addition, strong correlations between features may lead to redundancy, which reduces the efficiency of learning algorithms.

Dynamic Nature of Spam: The ever-changing tactics and techniques used by spammers, such as sophisticated language use and new structural manipulations, make maintaining a stable and effective feature set particularly challenging.

Computational Costs: Finding an optimal and distinct subset of features in high-dimensional spaces can be computationally intensive, requiring significant processing power and time.

Noise in Data: Features derived from long-term analysis of email content and metadata may include noise, blurring the distinction between spam and legitimate communications, complicating classification efforts.

Feature Subset Selection Using Metaheuristic Algorithms

To overcome the limitations associated with traditional feature engineering, researchers have gradually turned to metaheuristic algorithms for feature selection. These algorithms, inspired by natural and physical phenomena, excel at traversing complex, high-dimensional search spaces to identify the most valuable features (Talpur & O'Sullivan, 2020). By optimizing the selection process, they not only enhance classification accuracy but also reduce computational costs and improve the model's adaptability to sophisticated spam techniques. Some of the most prominent metaheuristic approaches used in spam detection include Wang & Zhao, 2021:

Genetic Algorithms (GA): It is based on the principles of evolution, using selection, genetic exchange, and mutation in an iterative manner to generate subsets of features that improve classification performance.

Particle Swarm Optimization (PSO): It is an algorithm that mimics social behaviors observed in swarms, such as those of birds or fish, directing candidate solutions toward optimal regions in the feature space by simulating movement and collective adaptation.

Ant Colony Optimization (ACO): It focuses on the foraging behaviors of ants, using pheromone trajectories to evaluate and optimize feature subsets. More stable trajectories encourage more efficient combinations.

Simulated Annealing (SA): When considering physical annealing in metallurgy, spectroscopy probabilistically adopts suboptimal solutions to circumvent locally optimal results, eventually progressing towards globally optimal subsets.

Harmony Search (HS): Leveraging the improvisational dynamics inherent in musical harmony, HS investigates the feature space by continuously changing and recombining features to obtain the most "consistent" subset for classification.

VII. RESEARCH DISCUSSIONS

The researchers' proposed spam detection model, which integrates natural language processing (NLP) with artificial neural networks (ANNs), achieved impressive results across various evaluation metrics. These results demonstrate the effectiveness of NLP-based preprocessing methods, such as TF-IDF, in identifying semantic and contextual features that distinguish spam emails from genuine ones. When combined with ANNs, these features resulted in improved accuracy, precision, and recall, confirming the method's suitability for contemporary spam filtering systems. Compared to traditional classifiers such as naive Bayes and support vector machines, the ANN-based model demonstrated superior performance in capturing complex, nonlinear relationships within the dataset. These results are consistent with recent research highlighting the benefits of hybrid and deep learning methodologies in addressing the increasing complexity of spam content. This analysis highlights the importance of integrating advanced machine learning techniques to enhance detection effectiveness. The implications of these findings go beyond technical precision. Effective spam detection not only improves user experience by alleviating inbox congestion but also enhances cybersecurity by reducing the risks associated with phishing and malicious communications. Given that email remains a primary channel for personal, professional, and financial interactions, implementing intelligent filtering systems is critical to protecting digital communications. However, some limitations were noted. The size of the dataset and its inherent imbalance posed challenges that necessitated extensive preprocessing, including resampling methods. Furthermore, the artificial neural network (ANN) model exhibited significant computational requirements, which may limit its scalability in real-time or resource-constrained environments. Another challenge is the linguistic diversity of spam, which includes the use of different languages, slang expressions, and deliberate misspellings. This diversity can reduce the generalizability of models trained on monolingual or uniform datasets. Looking ahead, several potential directions for future research are evident. Incorporating more advanced deep learning, such as Transformers and BERT, has the potential to improve the model's ability to understand contextual cues and subtle language patterns. Furthermore, expanding the dataset to include a variety of multilingual spam examples would enhance both robustness and generalizability. Additionally, the use of hybrid ensemble methods that combine artificial neural networks (ANNs) with classifiers such as decision trees or support vector machines (SVMs) would enhance their resistance to advanced spam techniques. Ultimately, achieving a better balance

between accuracy and computational efficiency will be essential for implementing spam detection systems in real-time scenarios.

Table 4: Comparative Analysis Of Spam Detection Techniques

Technique	Description	Key Strengths	Limitations / Challenges	Performance Insights
NLP + ANN (Proposed)	TF-IDF + ANN hybrid	High accuracy Robust semantic feature extraction	High computational cost Sensitive to linguistic diversity and class imbalance	Outperforms all other models Fast in detecting sophisticated spam
ANN Only	ANN on raw email features	Captures non-linear patterns	Lower semantic understanding	Good, but less effective than a hybrid
Naive Bayes	Probabilistic word-based	Fast and simple Interpretable	Assumes feature independence	Better performance than baseline, lower than ANN
SVM	Margin-based classifier	Handles high-dimensional data	Computationally intensive	Better than Naive Bayes, lower than hybrid
TF-IDF + Classical ML	Feature extraction + ML	Captures term importance	Limited pattern modelling	Baseline, less robustness

Table 4 presents a structured comparison of several spam detection techniques, highlighting their descriptions, strengths, limitations, and overall performance. Among the evaluated approaches, the proposed NLP + ANN hybrid model clearly demonstrates superior performance. By integrating TF-IDF–based feature extraction with an artificial neural network, this approach achieves higher accuracy and more robust semantic representation, enabling it to effectively capture complex linguistic patterns commonly found in sophisticated spam messages. Although the hybrid model involves higher computational cost and sensitivity to linguistic diversity and class imbalance, its ability to outperform all other techniques justifies this trade-off. In contrast, models such as ANN only, Naive Bayes, and SVM show reasonable performance but remain limited either by weaker semantic understanding or stronger modelling assumptions. The TF-IDF with the classical machine learning approach serves primarily as a baseline, offering limited robustness. Overall, the results indicate that combining advanced NLP techniques with neural networks provides the most effective and reliable solution for spam detection among the compared methods.

7.1 Ensemble Learning Techniques and Ensemble Feature Selection

Ensemble learning has become a powerful approach in machine learning, especially in spam detection systems, due to its ability to combine different models to achieve improved predictive performance compared to standalone classifiers. By leveraging the complementary strengths of various models or feature selection methods, ensemble techniques enhance both the robustness and the accuracy of spam detection, mitigating the weaknesses inherent in single approaches.

Ensemble Feature Selection

Feature selection plays a pivotal role in optimizing spam detection systems, and ensemble feature selection extends this by integrating the outcomes of multiple feature selection techniques. This approach aggregates diverse perspectives on feature importance, ensuring that the selected subset of features is not biased by the limitations of any single method. Several ensemble feature selection techniques are employed to address the complexities of high-dimensional spam datasets:

Voting-Based Methods: These combine feature classification methods produced by different algorithms, and the final subset is determined based on majority voting or weighted evaluation. This consensus-based approach reduces the risk of excluding important features.

Bagging: Also known as pre-pooling, pooling applies feature selection to multiple resampled subsets of data where the pooling of selected features from these subsets improves the stability and generalizability of the selected feature set.

Boosting: Iterative boosting emphasizes features that are difficult to capture, improving the selection process by adjusting weights to prioritize aspects that were previously underrepresented or misclassified.

Stacking: This approach combines the outputs of several feature selectors and feeds them into a comprehensive model that determines the final subset of features. By

leveraging the combined outputs, the clustering provides an accurate and efficient feature selection process.

Of these methods, combined attribute selection ensures a more comprehensive and balanced identification of relevant attributes, effectively reducing the risk of omitting important attributes or adding redundant attributes. This is critical in spam detection, where subtle textual and behavioral cues often determine classification success.

7.1 Classifier Ensemble Learning in Spam Detection Systems

Classifier ensemble techniques are similarly instrumental in enhancing spam detection performance. By combining predictions from different classifiers, ensemble techniques reduce classification errors and enhance their resilience against the ever-changing tactics used by spammers. Some of the most prominent ensemble techniques used in learning classifiers include:

Bagging (Bootstrap Aggregating): This technique involves training multiple classifiers on modified subsets of the dataset and then averaging or voting on their predictions. Aggregation effectively reduces variance and mitigates overfitting, making it particularly effective for noisy spam data.

Boosting: Unlike aggregation, boosting trains classifiers sequentially, focusing on cases misclassified by previous models. This iterative reweighting method enhances the classifier's ability to identify hard-to-detect spam.

Stacking: Aggregation combines predictions from various base classifiers with a super classifier, which learns the optimal way to combine these outputs to produce the best final predictions. This multi-layered strategy enhances generalization across a variety of spam patterns.

Random Forests: It is a specific type of ensemble that generates an ensemble of decision trees, each trained on random subsets of data and features. This approach captures complex nonlinear relationships within spam data, improving detection accuracy. Using these clustering techniques, spam detection systems improve their ability to adapt to sophisticated spammers' tactics, harnessing the diversity of multiple learning methods to build more robust and reliable defenses.

Several previous studies have played a significant role in developing spam detection systems using various methodologies. Korintzis et al. (2022) demonstrated the effectiveness of hybrid models that integrate deep learning with natural language processing, resulting in significant improvements in detection accuracy, albeit at the expense of high computational requirements. Wang and Zhao (2021) conducted a comprehensive review of both traditional and machine learning-based approaches, highlighting key challenges such as the dynamic nature of spam and dataset limitations. However, they did not provide empirical evidence for the newly proposed methods. Al-Jaradi et al. (2016) conducted a systematic review focusing on metrics and models for phishing detection, emphasizing the importance of machine learning and heuristics, but they did not directly address spam detection. Abbasi and Chen (2008) evaluated a set of text classification algorithms for spam detection, providing valuable insights into algorithm selection, although their research did not delve into deep learning techniques. Liu and Huang (2020) presented an ensemble learning framework for spam detection in public email systems, which improves robustness and accuracy compared to standalone models but requires significant computational resources. Finally, the paper "Spam Detection Based on Enhanced Natural Language Processing (NLP) and Deep Neural Networks (DNNs)" emphasized that combining enhanced natural language processing (ANN) with natural language processing can reduce false positives and improve classification performance, further enhancing this research.

VIII. CONCLUSION

This research focuses on spam and the significant threat it poses to communications, compromising user privacy and security. Technological advancements have weakened traditional filtering methods, such as blacklists, traditional machine learning classifiers, and rule-based approaches. This research focuses on providing a systematic analysis of the techniques and models used in spam detection. It also examines how these models, such as natural language processing (NLP) and artificial neural networks (ANNs), extract text features using TF-IDF and classify them to capture complex and nonlinear patterns in email data. The research relied on selecting twenty relevant, recently published studies from more than 80 academic references based on their methodological rigor, direct relevance to spam filtering and artificial intelligence, and their contributions to this field. These studies were then analyzed and compared, focusing on the techniques and evaluation strategies used in each. The analysis in this research encompassed a range of methodologies, ranging from classical machine learning algorithms (naive Bayes and support vector machines) to machine learning, deep learning, and hybrid models. The challenges and shortcomings of traditional methods, such as blacklists, content-based filtering, and rules-based approaches, due to contemporary spam campaigns, were also discussed.

This research emphasizes the effectiveness of integrating natural language processing (NLP) techniques with artificial neural networks (ANNs) as a comprehensive framework for spam detection. The findings indicate that NLP-driven preprocessing

methods, particularly TF-IDF, effectively extracted both semantic and contextual features. The integration of nonlinear learning capabilities of ANNs resulted in significant improvements in accuracy, precision, and recall. These results confirm the suitability of this strategy for modern spam filtering systems that require robustness and adaptability. Compared to traditional classifiers such as Naïve Bayes and SVM, the proposed artificial neural network-based model demonstrated improved performance in managing complex and advanced spam patterns. This aligns with recent research highlighting the benefits of hybrid and deep learning approaches in addressing the increasing complexity of spam content. In addition to technical accuracy, these results hold greater significance for cybersecurity, as effective spam detection mitigates malicious threats, including phishing attacks, while simultaneously improving the user experience. However, it is important to recognize some limitations. The size of the dataset and the current imbalance between classes posed challenges that required extensive preprocessing and resampling techniques. Furthermore, the high computational cost associated with ANN raises concerns about scalability in real-time or resource-constrained environments. Another obstacle is the linguistic diversity of spam, which includes multilingual texts, informal language, and obfuscation techniques, which may limit the applicability of developed models to standard or monolingual datasets. Therefore, future research should prioritize the integration of advanced deep learning frameworks, such as Transformers and BERT, to improve contextual understanding and semantic representation. Expanding datasets to include a variety of multilingual spam examples will be essential to enhance robustness and generalizability across different languages. Furthermore, the study of hybrid ensemble approaches that combine artificial neural networks with complementary classifiers (such as decision trees or support vector machines) may enhance the ability to adapt to complex spam techniques. Ultimately, achieving the optimal balance between accuracy and computational efficiency is critical for the effective implementation of spam detection models in real-time applications.

Authors' Contributions

Afaf Abdu Yahya Al-Nowidi: Principal Author.
Abdulaziz Ahmed Thawaba: Principal Author.
Maqbol Ahmed: Reviewer.

Conflict of Interest

The authors declare that there is no conflict of interest.

REFERENCES

1. Abbasi, A., & Chen, H. (2008). Comparative study of text classification techniques for spam detection. *Decision Support Systems*, *44*(1), 210–223.

2. Abbasi, M. S. (2023). *Automating behavior-based ransomware analysis, detection, and classification using machine learning* [PhD Thesis, Open Access Te Herenga Waka-Victoria University of Wellington]. https://openaccess.wgtn.ac.nz/articles/thesis/Automating_Behavior-based_Ransomware_Analysis_Detection_and_Classification_Using_Machine_Learning/22180858?file=39410965

3. Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H. A., & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *Ieee Access*, *7*, 70701–70718.

4. Al-Garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). A systematic review of phishing detection methods, metrics, and models. *Computers & Security*, *57*, 79–97.

5. Ali, W. N. H. W., Mohd, M., Fauzi, F., Shirai, K., & Noor, M. J. M. (2021). Implementation of Hyperparameter optimisation and over-sampling in detecting cyberbullying using machine learning approach. *Malaysian Journal of Computer Science*, 78–100.

6. Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: Review, approaches and open research problems. *Heliyon*, *5*(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

7. Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), 1.

8. Kourentzes, N., Petropoulos, F., & Trapero, J. (2022). Hybrid models combining deep learning and NLP for spam detection. *Journal of Intelligent Systems*, *31*(4), 451–470.

9. Liu, X., & Huang, J. (2020). Ensemble learning approach for spam detection in public email systems. *Expert Systems with Applications*, *143*, 113009.

10. Liu, Y., Zhou, Y., Yang, K., & Wang, X. (2023). Unsupervised deep learning for IoT time series.

IEEE Internet of Things Journal, *10*(16), 14285–14306.

11. Özbay, E. (2023). An active deep learning method for diabetic retinopathy detection in segmented fundus images using artificial bee colony algorithm. *Artificial Intelligence Review*, *56*(4), 3291–3318. https://doi.org/10.1007/s10462-022-10231-3

12. Rashid, Y., & Bhat, J. I. (2023). Topological to deep learning era for identifying influencers in online social networks :a systematic review. *Multimedia Tools and Applications*, *83*(5), 14671–14714. https://doi.org/10.1007/s11042-023-16002-8

13. Shehnepoor, S., Togneri, R., Liu, W., & Bennamoun, M. (2023). *Social Fraud Detection Review: Methods, Challenges and Analysis* (No. arXiv:2111.05645). arXiv. https://doi.org/10.48550/arXiv.2111.05645

14. Slodkowski, B. K., Da Silva, K. K. A., & Cazella, S. C. (2023). A systematic literature review on educational recommender systems for teaching and learning: Research trends, limitations and opportunities. *Education and Information Technologies*, *28*(3), 3289–3328. https://doi.org/10.1007/s10639-022-11341-9

15. Spark, A. (2023). *NLP + Multinomial Naive Bayes on Spark for spam detection*.

16. Talpur, B. A., & O'Sullivan, D. (2020). Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in twitter. *Informatics*, *7*(4), 52. https://www.mdpi.com/2227-9709/7/4/52

17. Text Mining + SVM + Decision Tree approach for spam detection. (2021). *Scientific Programming*.

18. Thakur, K., Ali, M. L., Obaidat, M. A., & Kamruzzaman, A. (2023). A systematic review on deep-learning-based phishing email detection. *Electronics*, *12*(21), 4545.

19. Wang, Y., & Zhao, L. (2021). Review of traditional and ML-based spam detection approaches. *IEEE Access*, *9*, 12254–12270.