AI-based Spam Detection Techniques for Online Social Networks: Challenges and Opportunities

Azza A. Abdo[1*], Khaznah Alhajri[2], Assail Alyami[3], Aljazi Alkhalaf[4], Bashayer Allail[5], Esra Alyami[6] and Hind Baaqeel[7]

[1*]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia. aaaali@iau.edu.sa,
Orcid: https://orcid.org/0000-0001-8246-9790

[2]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia. 2190004752@iau.edu.sa,
Orcid: Orcid: https://orcid.org/0000-0003-0987-8058

[3]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia. 2190005693@iau.edu.sa,
Orcid: https://orcid.org/0009-0005-7416-2293

[4]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia. 2190001702@iau.edu.sa,
Orcid: https://orcid.org/0000-0003-2586-836X

[5]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia. 2190004502@iau.edu.sa,
Orcid: https://orcid.org/0000-0002-5693-6356

[6]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia. 2190005556@iau.edu.sa,
Orcid: https://orcid.org/0000-0003-0395-1605

[7]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia. haaalssayed@iau.edu.sa,
Orcid: https://orcid.org/0000-0002-6343-6114

Received: June 10, 2023; Accepted: August 16, 2023; Published: August 30, 2023

Abstract

In recent years, online social networks (OSNs) have become a huge used platform for sharing activities, opinions, and advertisements. Spam content is considered one of the biggest threats in social networks. Spammers exploit OSNs for falsifying content as part of phishing, such as sharing forged advertisements, selling forged products, or sharing sexual words. Therefore, machine learning (ML) and deep learning (DL) techniques are the best methods for detecting phishing attacks and minimize their risk. This paper provides an overview of prior studies of OSNs spam detection modeling based on ML and DL techniques. The research papers are classified into three categories: the features used for prediction, the dataset size corresponding language used, real-time based

Journal of Internet Services and Information Security (JISIS), volume: 13, number: 3 (August), pp. 78-103.
DOI: *10.58346/JISIS.2023.I3.006*

*Corresponding author: Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, Saudi Arabia.

applications, and machine learning or deep learning techniques. Challenges and opportunities in phishing attacks prediction using ML and DL techniques are also concluded in our study.

Keywords: Spam Detection, Social Media, Twitter, Machine Learning.

1 Introduction

Online Social Networks (OSNs) have become an essential part of people's lives, allowing them to communicate and engage on a large scale. OSN refers to many information services that many people use (Barbier, G., 2011). The growing capabilities and popularity of OSNs have attracted many users. Twitter, for example, had 330 million monthly active users in 2021 (Delle, F.A., 2022). Twitter is a microblogging service that allows users to construct tweets (i.e., short messages). Tweets may include text, images, videos, or URLs. URL is commonly referred to as a website address on the World Wide Web (WWW). The increasing use of OSNs makes them an ideal source for data, and so have cybersecurity threats (Cortizo, J., 2009) (Alharbi, A., 2022). The COVID-19 pandemic has led to an overabundance of information on social media, with people sharing news, and opinions on a massive scale. However, not all information shared is accurate or misinformation related to COVID-19 (Mourad, A., 2020). Also, a face advertisement with phishing links was increased cording to scam the users. Cybersecurity threats include a wide range of potentially illegal behaviors on OSNs (Razzaq, A., 2013). The most widespread type of cybersecurity attack is phishing (Alharbi, A., 2022). Phishing is an online identity theft, which is one of the social engineering attacks in which the attacker tries to steal a user's personal information and sensitive data (Goel, D., 2018). Spammers rapidly exploit OSNs for falsifying content as a part of phishing. A spammer is a user who sends an enormous number of irrelevant contents to the receiver without the receiver's permission (Tandon, A., 2022). Furthermore, a recent report by the Anti-Phishing Working Group (APWG) shows that in the second quarter of 2022, many 1,097,811 phishing attacks were observed, which is the worst quarter that APWG has ever observed. Moreover, threats on OSNs continued to rise with a 47% increase from the first to the second quarter of 2022. Machine learning (ML) and data mining (DM) techniques (Goel, D., 2018), are currently employed to detect phishing attacks and spammer users. Machine learning is a scientific study of algorithms and statistical models used to utilize computers to operate human tasks. DM is the process of discovering previously unknown, valid patterns and correlations in a dataset using advanced data analysis techniques to increase the results accuracy of ML algorithms (Seifert, J.W., 2004). Some researchers apply ML to detect phishing attacks by depending on detect URLs malicious links spread through OSNs, which is known as the URL-based level. Other researchers focus on analyzing the spam content for the spammer's users over social media, which is known as the content-based level, and others depend on detecting the nature of the spammer account itself (account-based level), and the most recent researchers focus on using a hybrid of the three levels. At the phishing content detection level, natural language processing (NLP) was used for extracting a complete meaning from free text (Kao, A., 2007) (Pais, S., 2022), and it was applied to detect phishing content from a legitimate one, including machine translation and information extraction. However, datasets from OSNs are vast, noisy, and dynamic. Extracting useful information from OSNs data is only possible with DM (Baatarjav, E.A., 2008). With automated processing, OSNs data analysis becomes feasible in a reasonable amount of time. For example, highly insignificant tweets on Twitter make the data noisy. Furthermore, it is critical to consider frequent changes and updates over short intervals. DM can assist in overcoming the mentioned challenges and understanding data better to use them for research purposes (Baatarjav, E.A., 2008). Applying DM techniques to large OSN datasets has the potential to detect phishing attacks and spammers (Cortizo, J., 2009) (King, I., 2009). Additionally, there are three challenges discussed for detecting spam content in

social media: Language, dataset size, and time that the model takes for detecting the phishing content, and it affects the system deployment in real-time.

This paper presents an overview of related research on ML and DL techniques used to detect phishing attacks on OSNs. The main contributions of this paper are the following:

- Classify OSNs spam detection techniques-based ML and DL. Classification is based on several perspectives, including the features used for detection, dataset size, and real-time.
- Reviewing the previous research for OSN spam detection over the period 2010 to 2022, according to different spam types, and determining the ML algorithms used for detection.
- Emphasizing challenges and opportunities in OSNs spam detection.

The rest of the paper is structured as follows. Background about phishing attacks on social media are presented in section 2. Section 3 explains the literature selection methodology. Literature review papers are explained in detail in section 4. Discussion and results are shown in section 5.

2 Phishing Attack on Social Networks

Due to the significant financial gain and global publicity that OSNs, spammers use them and disseminate false and misleading information, so that users tricked illegally. On OSNs, there are different kinds of spammers (Kandasamy, K., 2014):

- **Phishers:** Spammers' accounts distribute malicious URLs in their tweets. When other users click these links, they lured to steal their personal information.
- **Malware**: Propagators tweet malicious links. When users click on the links, malware downloaded.
- **Marketing users:** Focus on spreading advertisements for products. This type of spammers is not harmful because their target is popularizing their business.
- **Adult content propagators spammer:** Spammers tweet adult content with an attached link, which redirects the user to a malicious website after clicking it.

To detect the phishing content on OSN, there are three primary factors that should be considered, URL attached into the social content, social content itself, the user account who shares the content. Each of the factors explained as the following:

1. **URL attached into the social content:** URL is the widely attack approaches for phishing attacks. Therefore, the components of URLs should be understood to detect phishing attacks. There are different features to detect whether the URL is phishing or not, such as Internet Protocol address (IP), dots, and @ symbol. If there is an IP in the URL, dots are more than three, or if @ symbol is in the URL, it is considered phishing (Rahman, M.S., 2021).

Figure 1: URL Components

Figure 1 shows the basic structure of URLs. A URL in its standard form starts with the protocol name used to access the webpage. Here, "https" is the protocol name. Then, the domain name (i.e., hostname) of the webpage. The domain name contains multiple parts: the subdomain, the Second Level Domain (SLD), and the Top-Level Domain (TLD). The subdomain is the part preceding the second-level domain, which corresponds to the organization name in the host server. TLD shows the domains in the Internet Domain Name System (DNS) root zone. Finally, the page's path indicates the inner

address. Some parts of the URL can be easily found or bought for phishing, such as SLD, which only shows the organization's (i.e., URL owner) name. In addition, because the inner address structure is directly dependent on the owner, an attacker can generate an infinite number of URLs by extending the SLD with path and file names. However, the domain name is the unique and crucial part of the URL, which consists of SLD and TLD (Sahingoz, O.K., 2019). According to (Sahingoz, O.K., 2019), URL phishing attacks can be generated using one or more of the following tricks:

- **Cybersquatting:** Registering a trademark (i.e., owned mark) as a domain name without having a legitimate claim to use it (Gilwit, D.B., 2003).
- **Typo squatting (URL hijacking or fake URLs):** Typosquatting is a type of cybersquatting where the hacker takes advantage of the incorrect typing of URL to direct users to malicious sites, such as using "exmaple.com" instead of "example.com" (Sahingoz, O.K., 2019).
- **Dotted decimal IP address:** Using the IP address directly in the URL or the anchor text instead of the DNS name (e.g., http://209.191.122.70/)
- **Special characters:** Indicates that all text before the symbol is a comment (e.g., @ in the visible link).
- **Random characters:** Using a too long URL with an overlay of random characters to obscure the true domain and try to hide the location of the malicious file, such as "http://innocent.com/irs.gov/logn/fasdkf.sdrfgvgh.hfdgdjhgdftghd/adfgjgjhgjffhgfhj/ght.php".
- **Combined word usage**: Using a combination of two or more different words in a meaningful order to make a website appear legitimate to users (Sahingoz, O.K., 2019).

2. **Social content text:** Capturing all the semantic properties extracted from the text of tweets, including different attributes such as the length of a tweet, hashtag count, and the number of spam words in the tweet (Wang, B., 2015). Natural Language Processing (NLP) is an AI approach used to detect spam content in social content by analysing texts. NLP is concerned with enabling computers to analyse and understand the original text or human speech to perform the desired tasks. NLP is a combination of rule-based modelling of human language with statistical machine learning models, deep learning (DL) models, and computational linguistics combined together, allowing computers to process human language involving text data and understand it, including the writer's intent and sentiment (Marie-Sainte, S.L., 2018).
3. **Social user account:** Refer to features generated from particular characteristics of user behaviour, the main features used are the number of followings, followers, and reputation (Ho, K., 2018).

3 Literature Selection Methodology

Literature selection methodology Extensive literature exists in the field of phishing in social media. In this paper, we concentrated on the literature of the period 2010 to 2022. The search results were filtered into three filtering phases, keywords, title, and abstract. The literature selection methodology phases are summarized in **Figure 2**.

Figure 2: Literature Selection Methodology Phases

In the keywords filtering phase, the selection methodology of related work began with a search from the publishers' online search engines using different related keywords, including spam detection using ML, phishing detection on OSNs, spam detection on Twitter, phishing detection techniques, and Arabic spam/phishing detection. This phase resulted in 1446 related work: 514 IEEE articles, 602 Springer articles, 198 ACM articles,56 MDPI, and 132 Elsevier articles. Based on the result of the previous phase, the titles of the articles were reviewed and the most relevant articles were selected. This phase resulted in 120 related works: 32 IEEE, 55 Springer, 21 ACM, 20 MDPI, and 12 Elsevier articles. In the abstract filtering phase, the abstract reading was conducted for the resulting articles from the previous phase. Based on those readings, the most relevant articles were selected. This phase resulted in forty-five related works: 16 from IEEE, 12 from Springer, 6 from ACM, 5 from MDPI, and 6 from Elsevier. The percentage of articles per publisher in this filtering phase is illustrated in **Figure 3**.

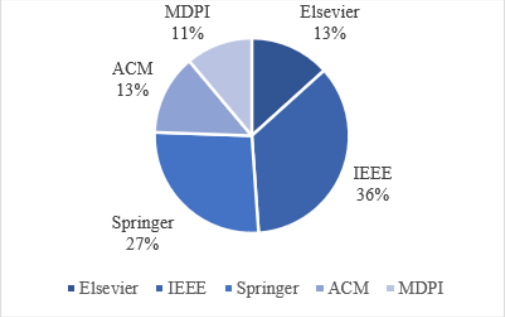

Figure 3: Percentage of Articles Per Publisher

4 Literature Review

Taxonomy of Literature Reviews

The related research taxonomy is divided into six categories, as shown in **Figure 4** below. The reviewed study articles are categorized as follows:(1) feature types for detection. (2) whether the model was deployed and used in real-time or not, and (3) The size of the dataset related to the language of the dataset.

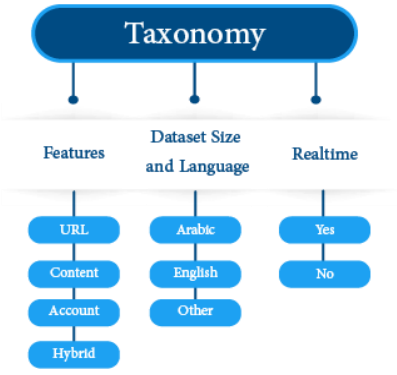

Figure 4: Taxonomy of the Related Work on Phishing Attacks Detection

Based on Features

In this paper, forty research on detecting phishing attacks are reviewed and classified based on four types of features selected, which are:

- **Web pages and their URL-based Features:** Features related to URLs, refer to phishing webpages. These URLs may be included with mobile messages, Emails, or tweet content.
- **Content-based Features:** Features related to the content of email, webpage, or tweet.
- **Account-based Feature:** Features related to the user's account, such as network (i.e., graph-based) and profile features.
- **Hybrid-based features:** Such as account and content-based features, for example, features related to the user's account and the content of a tweet or post. Additionally,
- **URL and other based features**, such as account, content, and forwarding-based.

a) Web Pages and their URL Based Features

In (Sahingoz, O.K., 2019), real-time detection for phishing web pages was analyzed. the authors used the URL phishing detection system using ML algorithms to detect the phishing pages. Each URL was analyzed using Natural Language Processing based Features (NLP). 40 different NLP-based features were extracted using NLP for URL, and 1,701 word-features are extracted from the URL dataset. The total features used for the hybrid was 1,741 and then decreased to 140 features. The dataset size was a total of 73,3575 URLs, 36,400 legitimate, and 37,175 phishing. The system used seven machine learning algorithms, which are Naive Base (NB), Adaboost, K-star (n=3), Random Forest (RF), Sequential Minimal Optimization (SMO), Kstar, and Decision Tree (DT). A comparison between the seven models is applied after training models into NLP-based URL features, URLs word vectors, and a hybrid of the two features, the results show that the hybrid of URL features and word vector increased the performance at the rate of 2.24% according to NLP based features and 13.14% according to word vectors. The use of NLP features in (Sahingoz, O.K., 2019) has a significant effect on the accuracy results. Also, huge size of phishing and legitimate data, real-time execution, independence from third-party services, and use of feature-rich classifiers, are strengths of the detection model in (Sahingoz, O.K., 2019).

In (Xiang, G., 2011), the authors proposed an ML anti-phishing solution for websites based on HTML and URL features. The proposed model was called as named CANTINA+. The system exploits HTML Document Object Model (DOM) detection by three major phases. In the first phase, a hashing filter for highly similar phish examined a Web page to get the similarity with known phishing attacks. Secondly, the system checks if the web page uses login forms to request sensitive information. Thirdly, utilizing 15 features with machine learning models to classify Web pages, in which Bayesian Network (BN) ML algorithm was used. CANTINA+ was evaluated under using with 8118 phishes and 4883 legitimate Web pages, and achieved more than 99% TP, with F1=0.9894. Algorithm agnostic and comprehensive layered approach are strengths of the detection system in (Xiang, G., 2011). On the other side, the model is unable deal with Cross Site Scripting (XSS) and phishing web pages purely made up of images, leaving the algorithm with no text to analyze.

In (Lakshmi, V.S., 2012), the authors suggested a detection model using ML algorithms to detect malicious websites based on URL features. 17 features related to URL and HTML source code were used. PhishTank was used to collect 100 malicious and 100 legitimate websites for model training. Three ML classification algorithms were used, which are Multi-Layer Perception (MLP), Decision Tree Induction (J48), and NB. The classifiers were evaluated by a ten-fold cross-validation algorithm and two performance criteria: Detection accuracy and training time. By comparing the three algorithms' results, Decision tree Induction (J48) outperformed other algorithms with an accuracy of 98.5%.

In (Burnap, P., 2015), the authors developed a real-time system to determine the malicious URLs on Twitter based on ML. Machine activity log data features such as CPU usage, network traffic, and network connection statistics after linking streamed Twitter data to a high interaction honeypotNB. The

dataset size was 130,503 tweets, 122,542 of which contain URLs. 2000 tweets from each dataset were randomly selected. BayesNet (BN), NB, DT, and MLP algorithms were used to determine the best algorithm based on accuracy. During the training, DT and MLP algorithms achieved up to 97% accuracy. The detection method utilized yielded promising results in predicting malicious behavior within seconds of URL interaction. However, its limited generality to other events and the challenge of generalization to unseen data should be further explored.

In (Jain, A.K., 2018), the authors proposed a website anti-phishing system using ML algorithms, called PHISH-SAFE. The dataset was composed of 32,951 phishing URLs, and 2500 legitimate URLs downloaded from different public sources. Two ML algorithms were used for training and testing, which are Support Vector Machine (SVM) and Naïve Bayes (NB). The used technique leverages a substantial dataset of phishing and legitimate URLs, employs 14 distinct features for feature extraction, and conducts thorough experiments using different classifier sizes, resulting in high accuracy, exceeding 90% with the SVM classifier. However, it has some weaknesses, including the limited size and potential bias in the non-phishing dataset, the absence of validation on real-world data, and the need for more in-depth study of false positives and false negatives.

In (Shivangi, S., 2018), the authors proposed a DL method to detect malicious URLs on Twitter. The used algorithms were Dense Artificial Neural Network (ANN) and Long-Short-Term Memory (LSTM). Five different datasets to test the algorithms in the models were used: 512, 2343, 11234, 45630, and 456300. As a result, ANN indicates higher accuracy than LSTM if the data set size is smaller, while if the data set size is significant, the LSTM has better results with an accuracy of 96.89%. Additionally, the model was deployed as a real-time Chrome extension. The model meets various non-functional requirements, such as efficiency, ease of deployment, and reduced training time, making it suitable for real-world applications. However, weaknesses lie in the LSTM model's potential overfitting on smaller datasets and the dependency of ANN accuracy on the size of training data.

In (Liew, S.W., 2019) the authors proposed a security alert mechanism to detect phishing tweets in real-time using ML algorithms. To design this mechanism, the authors divided the work into two stages: Training and formulation. At the training stage, Random Forest (RF) model was used on a dataset containing 2973 samples. While in the formulation stage, 11 features embedded into the system. Moreover, cross-validation of ten folds was used for testing. Furthermore, the system tested 200 phishing URLs posted on Twitter and got 195 out of 200 URLs as phishing by displaying a red indicator next to the tweet, concluding that the system’s accuracy reached 94.75%. The technique's generalizability to multiple datasets and platforms is uncertain, and no comparison to other phishing detection approaches was offered.

b) Content-based Features

In (Yao, J., 2022), the authors presented a hybrid BiGRU-CNN network with joint textual and phonetic embedding for detecting spam in the Chinese language. Due to the extremely common phenomena of homophony in Chinese, textual and phonetic embedding were used throughout the article for the identification of spam in that language. When training for word embedding, phonetic information cannot be ignored. When utilized in Chinese spam detection tasks, two methods of joint textual and phonetic embedding are being investigated by the authors to capture fuller representations of Chinese text. BiGRU-CNN-JE model in (Yao, J., 2022) achieved over 0.94 accuracy. The used technique exhibited notable strengths in in enhancing spam detection on a Chinese chat dataset. However, the method's limited capability to handle other types of textual noise and its increased time complexity due to additional model parameters pose challenges. Further research is necessary to address these issues.

The authors of (Zhao, C., 2020) concentrated on addressing the class imbalance problem in spam detection when spam messages outnumber genuine communications in social networks. They suggested a heterogeneous stacking-based ensemble learning technique for balancing training between base classifiers and metaclassifiers at the algorithmic and data levels. Their solution consists of a two-level structure with a base module and a combining module, which allows for the use of multiple learning approaches as basis classifiers to increase learning impact. Furthermore, for the ensemble, they used a cost-sensitive learning-enhanced deep neural network to offset the influence of uneven class distributions on classification performance. Their trials were carried out on a dataset of 600 million tweets, of which 6.5 million were identified as malicious. The suggested approach outperformed standard machine learning techniques on the same dataset, achieving an F1-score of 70%. Overall, the method provides a strong spam detection framework for social networks, successfully correcting class imbalance via the ensemble approach. To improve the method's performance, new efforts are planned to investigate deeper hidden feature representations and test classifiers with alternative dataset features.

In (Sharmin, S., 2017), the authors proposed an ML technique that analyzes the comments on YouTube videos and classifies them as spam or not. Five datasets composed of 1,956 real messages from five video comments on YouTube were used. Additionally, Five ML algorithms were used: NB, 1-KNN, 3-KNN (K-nearest neighbors), Bagging, and SVM. The aim was to find the best classifier in accuracy, recall, classification error, F-measure, and Matthews Correlation Coefficient (MCC). As a result, the NB and Bagging have an accuracy average above 90% in four datasets. However, the NB result was less than 90% on one data set of one channel. While the technique shows potential, it does have significant drawbacks, including a limited scope to YouTube comments and a lack of extensive dataset and assessment information. Future research should focus on expanding the method to additional social media datasets, identifying spam account holders, and developing a real-time comment filtering plug-in for other platforms. Real-world validation and more detailed algorithm comparisons are required to increase its trustworthiness.

In (Wu, T., 2017), the authors proposed a DL spam detection model for Twitter. Four datasets were collected, with a total of 1,376,206 for spam tweets and 673,836 for nonspam tweets. The features used for comparison were content-based. Each dataset was split into 60% training data and 40% testing data for each run of tests. The proposed model was compared to different existing algorithms. According to the results, the proposed technique outperforms existing methods with an accuracy of 99.35%. The model is compared to different algorithms which help to decide the better model. However, the model comparison based only on one feature.

In (Sohrabi, M.K., 2018), the authors proposed an ML technique to detect spam comments on Facebook. The dataset was a total of 200,000 posts and comments collected in two ways: Via an agent for public pages and manually for personal pages. A feature selection approach was used to select 7 of the 13 content features based on their error rate. There were two proposed methods, a combination of clustering with SVM or DT. The clustering approach correctly detected 71.4% of spam messages and increased to 89.8% after combining it with SVM. However, the combination of clustering with the DT had an accuracy of 70.8%- and lower-time complexity than SVM. The technique for combining between two algorithms increased the result in significant way. Since the model collected data in two ways the dataset size considers small.

In (Kontsewaya, Y., 2021), the authors proposed an email spam detection model using ML algorithms with NLP techniques. The dataset used was about 5728 emails labeled as spam and legitimate from Kaggle. The dataset went through three main steps, which are analyzing, training, and testing. The dataset was analyzed using NLP and reduced to 5695 emails after the first step, with 4360 legitimate

emails and 1368 spam emails. Six ML algorithms were used in the training step, which are NB, KNN, SVM, Logistic regression (LR), DT, and RF. During the testing step, the algorithms were evaluated based on precision, recall, accuracy, F-measure, and Receiver Operating Characteristic (ROC) Area. The results showed that LR and NB achieved the highest accuracy, which reached 99% for spam detection. The model analysis the dataset to get the accurate content required for the training step. Also, many algorithms were used to get the better result. However, the dataset went through many steps which led to reduce the size of the dataset.

In (Koggalahewa, D., 2022), the research focused on identifying spam accounts on Twitter using ML algorithms based on user peer acceptability. The peer acceptance of another user was estimated using the two users' shared interests in various topics, which is a content-based feature. The researchers used three publicly available datasets with a total number of 5932 users. Using ML algorithms, the users were separated into 'focused' and 'diverse' groups. The peer acceptance is then evaluated to identify spammers. Spammers were detected with 96.9% accuracy. Different datasets help to improve the generality of the model because the datasets will cover many concepts.

In (Saeed, R.M., 2022), the authors presented an ML method for Arabic spam review detection. The method was tested on two datasets of varying sizes: 1600 and 94,052 using content-based features. Four experiments were carried out to evaluate the proposed method's performance. The results proved the approach's efficiency, with 95.25% classification accuracy. The model worked on Arabic content which is strong point since most of the research focused on English content. Using one feature sometimes may not be enough since they deal with Arabic content.

In (Alkadri, A.M., 2022), the authors analyzed Arabic spam content for Twitter OSNs, including advertising and malware. A collected dataset of Arabic spam that includes the original tweets and the related annotations. The authors used the approach of replacement embedding data augmentation for Arabic text to increase the size of the data. NB, LR, and SVM models were used for training machines on the dataset before and after the augmentation to compare the original (non-augmented) dataset results with the augmented data set. The best results were obtained for the augmented dataset. Using data augmentation help to solve the problem which is small size of the dataset.

In (Al-Azani, S., 2018), the authors used the ML method for Arabic spam detection on Twitter based on content (word embedding) features. Three algorithms were experimented with, namely: SVM, NB, and DT. For the dataset, it contained 1944 spam and 1559 legitimate tweets labeled manually. The results showed that SVM has the highest accuracy of 87.32% accuracy for the Arabic spam dataset model. The advantage of the detection model in (Al-Azani, S., 2018) that it is applied three different algorithms to define the best result.

In (Najadat, H., 2021), the authors proposed an ML detection model for Arabic spam in Facebook comments. The dataset was of size of 3,000 Arabic comments. The comments were classified using content-based features, such as the keywords extracted from them. Filter methods, term's weight, and Term Frequency–Inverse Document Frequency (TF-IDF) matrix were used to extract keywords from the comments. The used classifiers are DT, KNN, SVM, and NB. The DT classifier outperformed the other classifiers with an accuracy of 92.63%. Arabic content used in (Najadat, H., 2021) requires more than one feature to get better result, because Arabic content too complicated to deal with.

In (Najadat, H., 2021), the authors proposed an ML detection model for Arabic spam messages on OSNs that specifically targeted Saudi Arabia users. The detection system combined the Rule-Based scoring technique and NB classifier. A dataset of size 150 messages was collected from WhatsApp and SMS (50 spam and 100 non-spam). The dataset was manually labeled using content-based features. The

Rule-Based scoring technique achieved 52% accuracy, while the NB classifier achieved 86% accuracy. In (Najadat, H., 2021), the dataset covered specific culture which will led to reduce the generality of the model.

c) Account-based Features

This section presents the related work to phishing detection based on account features.

In (Alhassun, A.S., 2022), For identifying spam accounts on Twitter, the authors suggested a hybrid text and metadata-based deep-learning system. Two models were presented by the framework for detecting Arabic spam accounts on Twitter using deep convolutional neural networks: the first was based solely on text data, while the second incorporated text data and metadata from the tweets to fully utilize the data. The approach utilized two different forms of data: metadata and text-based data using a convolution neural networks (CNN) model. Out of the dataset, 12 features that were easy to calculate and extract were chosen. These features, including account age, follower count, and reply count, were taken from the user account and tweet data. The accuracy of the proposed framework, which came in at 94.27%, was the best in the combined model, demonstrating its supremacy. A total of 16,700 people were considered, and their tweets were created and classified. Combination between text data and metadata help to increase the generality of the detection model in (Alhassun, A.S., 2022), because it will cover many concepts.

In (Zhao, C., 2020), the authors proposed an attention-based graph neural network for spam bot detection in social networks. In this article, the authors discuss the increasing prevalence of spam bots and other malicious accounts on social media platforms and propose a novel approach for detecting and combating these issues. They learned an involved technique to combine the many neighborhood interactions between nodes to operate the directed social graph, using a graph neural network to build a detection model by aggregating features and neighbor relationships. The authors claimed that the PRAUC value for their suggested approach was 0.91. In (Zhao, C., 2020), they compared graphs for competing methods and found that graphs performed better in terms of accuracy and efficiency. Accordingly, the technique can detect spam activities in online social networks. Since only a node/vector dataset was tested, their approach was not generalizable to other datasets. Despite this, their experiment showed that it remains relatively stable.

In (Wang, A.H., 2010), the author used ML to identify spam accounts on Twitter. The used algorithms were DT, NN, SVM, and NB. For the dataset, around 25K users, 500K tweets, and 49M follower/friend relationships in total were collected. The main finding of this research is that the reputation feature among graph-based systems has the best performance for detecting spam accounts. The results showed that the NB algorithm achieved the best performance with 89% precision. This study had the advantage that it included a large dataset. Also, this study offers valuable insight into the effectiveness of ML for spam detection on Twitter.

In (Alharthi, R., 2019), the authors proposed an ML detection model for Arabic spammer accounts on Twitter. The dataset was composed of users' and spammers' groups (i.e., spam groups or promote groups) accounts. The authors suggested 16 features related to users' behavior. Moreover, label propagation and label spreading algorithms were used. The model achieved an accuracy of 91%. Additionally, to evaluate the model, 20 accounts were manually selected, and compared to the detected result. Results showed that 18 accounts were identified successfully and two failed due to the similarity in users' behavior. A thorough analysis of the model's strengths revealed that it could detect malicious accounts with high accuracy.

In (Chy, M.K.A., 2019), the authors suggested an ML phishing detection model using clustering techniques. A closed-ended questionnaire was used to collect the dataset in 60 days, and several 1,000 Facebook users' data was collected. Additionally, phishing has affected 50.5% of users. Six classification algorithms were used, which are: RF, SVM, NB, Neural Network (NN), DT, and LR. As a result, LR has promising findings from many parameters with an accuracy of 99.8% accuracy. The authors demonstrated that their study met the objective. However, the data that is used was insufficient.

In (Wei, F., 2019), the authors proposed a DL model to differentiate Twitter bots from genuine accounts. The model employed recurrent neural networks, specifically bidirectional Long Short-term Memory (BiLSTM) with word embedding to efficiently capture features across tweets. The dataset considered 3,474 genuine accounts with 8.4 million tweets and 1,455 bot accounts with 3 million tweets. The model achieved an accuracy of 96.2%, which is a similar performance compared with the presented existing work. The robustness of the model was also demonstrated by its ability to identify new social bots that had not been seen before.

In (Elyusufi, Y., 2019), authors proposed a fake Facebook profile detection approach based on ML. The authors used 2816 profiles, including fake and legitimate profiles, 80% of the dataset was used for training and 20% for testing. Two algorithms were used: DT (J48) and NB. The authors selected 33 profile features in the first feature selection phase. Then they observed unneeded features. To make the model more effective, they reduced the number of features to 4. The algorithms were compared to identify the most effective classifiers. The results showed that the DT (J48) algorithm performance was better than the NB algorithm with an accuracy of 99.28%. This proposed model used a limited number of users profiles.

In (Sowmya, P., 2020), the author proposed a detection method for fake and cloned accounts on Twitter using ML. The central architecture is applied based on the number of modules, and the username or the avatar can detect fake profiles. Accounts are made by lots of tweets without their location or not having any tweets yet. The dataset contained 1100 fake accounts, 1100 genuine accounts, and 800 cloned accounts. Two approaches were used, similarity measures and the C4.5 algorithm. The findings of fake account detection are set with an accuracy of 90.20%. Moreover, the results of cloned accounts detection were obtained with an accuracy of 90% using both similarity measures and the C4.5 algorithm.

In (Bharti, K.K., 2021), the authors proposed a detection technique that uses ML algorithms to detect accounts with fake Twitter followers based on account features. The dataset contained a total of 6973 accounts data from different sources. To classify an account as legitimate or fake, LR integrated with Particle Swarm Optimization (PSO) was used. The proposed model is then compared to the competitive state-of-the-art competitors, such as NB, DT, and LR. The findings demonstrate that the proposed model performs better performance than others in many cases, which correctly classified 79.9% of malicious accounts. For (Bharti, K.K., 2021), Collecting data from different sources helpel to increase the generality of the model. However, one feature to detect fake accounts may not enough for accurate result.

d) Hybrid-based Features

This section presents the related work to phishing detection based on hybrid features.

In (Zheng, X., 2016), the authors proposed an ML model for spam account detection on Sina Weibo. The dataset size was about 25,000, almost the last 500 messages for 50 users. The dataset was manually classified into spammer and non-spammer categories. A set of 18 features is then extracted from contents and user account activity. The used algorithm is Extreme Learning Machine (ELM-based). The results showed that the model successfully identified 99.1% of spammers and 99.9% of non-spammers. The

authors then compared their model against SVM-based algorithms (DT, NB, and BN). Both ELM and SVM classifiers obtain excellent accuracy in the comparison, but the ELM-based algorithm was more efficient due to its speed.

In (Mataoui, M.H., 2017), the authors proposed a spam detection system using ML algorithms by NLP of Arabic content on Facebook. The dataset was collected from Facebook with a total of 99 posts and 9697 related comments. The dataset was manually analyzed as spam or non-spam content. Nine content-based and account-based features were extracted. Seven classifiers were used, namely: NB, J48, Sequential minimal optimization (SMO), Decision Table, LR, and Locally Weighted Learning (LWL). The algorithms were evaluated by a ten-fold cross-validation The results showed that the J48 outperforms other algorithms with 91.73% of correctly classified instances for the unbalanced dataset and 76.57% of correctly classified instances for the balanced dataset. The weakness of this model was the low performance on the balanced dataset.

In (Ho, K., 2018), the authors proposed an ML spam detection model named WEST (Workbench Evaluation Spammer Detection system in Twitter) on Twitter using content and account-based features. The authors collected 1729 tweets for the dataset, 206 as spam and 1528 as legitime. Five different algorithms were used to compare results: SVM, DT, NB, KNN, and RF. The experiment results show that the most effective and efficient selection of features for detecting spammers are the time-related activity feature and the tweet content feature. The number of legitimate tweets is much greater than the spam tweets that could have bias result.

In (Alorini, D., 2019), the authors proposed an ML model to detect spam tweets based on account and content features in the Gulf Arab region on Twitter Arabic hashtags. The dataset crawled from Twitter with a size of 2000 tweets. The used ML algorithms were NB and SVM. The results show that as the number of tested tweets increases, the accuracy decreases. However, the NB algorithm produced more accurate results for detecting spam tweets with an accuracy of 86%. The authors mentioned that the limitation of the study is the small size of the dataset which may not be representative of the entire population of tweets in the Gulf Arab region. The strength of the study is that it was conducted in the context of the Gulf Arab region which contributes to a better understanding of the region's online culture.

In (Jose, T., 2019), the authors proposed a spam detection model based on ML to discover spammers that appear to produce legitimate tweets on the Weibo Chinese Twitter platform. The dataset collected focuses on detecting fake followers and user interest topics, it means the account and content-based features. A technique named Latent Dirichlet Allocation LDA topic modeling was utilized. Precision, recall, and F1-score are the three measures to assess four classification quality, namely: SVM, NB, DT, NN. The proposed technique surpasses other state-of-the-art approaches in terms of the average F1 score.

In (Barushka, A., 2020), the authors proposed a DL model to detect spam on two OSNs, which are Hyves and Twitter using MOEFS and RDNN algorithms with bagging. The datasets were crawled from both Hyves and Twitter, with 821 messages and 61,675 tweets, respectively. The extracted features were based on content and account features. Ten-fold cross-validation was performed and several methods used. RDNN, NB, SVM, AdaBoost M1, and RF were compared. The proposed approach achieved an accuracy of 90.02% and 95.60% for Hyves and Twitter, respectively. The number of Hyves's extracted messages is much less than the Twitter's tweets, which could have biased results.

In (Mubarak, H., 2020), the authors presented an ML spam detection model on Twitter. The model was designed by analyzing Arabic spam tweets and accounts. The dataset was composed of 134,222 tweets collected from the MENA (Middle East North Africa) region. The research concentrates on content and account-based features. State-of-the-art classification techniques were used, including

Arabic Bidirectional Encoder Representation from Transformers (AraBERT). The findings demonstrated that both SVMs combined with n-grams and AraBERT could recognize spam tweets with 99.4 and 99.7 accuracies, respectively. For spam accounts, the results showed that the tool had difficulties in recognizing spam accounts when they post tweets that toggle between Arabic and English.

In (El-Mawass, N., 2016), the authors used ML method for Arabic spammers detection on Twitter based on account and content features. Three algorithms were used, namely: NB, RF, and SVM. A sample of 5000 tweets out of more than 23 million Arabic tweets were manually labeled by the authors. The results showed that RF has the highest accuracy of 92.59% for the Arabic spam detection model. Using Arabic dataset could be one of the challenges for this paper which make it harder to generalize the model for other OSN

In (Kandasamy, K., 2014), the authors proposed an ML detection approach to detect spam accounts on Twitter. The integrated approach comprises the use of URL analysis, NLP, and ML techniques. The dataset contained 10 recent Tweets from 100 users. Six URL and content-based features were used for classification. The approach was compared to two ML algorithms, which are NB and SVM. The proposed approach outperformed the mentioned algorithms used alone with an accuracy of 98%.

In (Cao, J., 2016), the authors developed a model to detect malicious URLs on a Chinese OSN called Seina Weibo based on URL, forwarding, and graph-based features. The dataset was crawled from Sina Weibo with a size of approximately 100,000 messages. The used algorithms were BN, J48, and RF. Ten-fold cross-validation within ML methods was used. The main finding is that forwarding-based features are much more effective than other features and Blacklists, with an average accuracy of 80%. Using dataset from Weibo could be one of the challenges for this paper which make it harder to generalize the model for other OSN.

In (Aggarwal, A., 2012), the authors developed a real-time phishing detection system on Twitter called PhishAri as a Chrome extension. Blacklists and ML algorithms were used to detect phishing based on various features. 23 features were used and divided into four categories: URL-based, Tweet-based, Network-based, and WHOIS-based. The dataset size was 309,321 tweets, filtering out tweets with URLs. Three ML algorithms were used, NB, DT, and RF. As a result, the RF classifier performs best with an accuracy of 92.52%. The researchers demonstrated that their system outperforms standard blacklisting mechanisms like PhishTank, Google Safebrowsing, and Twitter's defense mechanism.

In (Djaballah, K.A., 2020), the authors developed an application based on ML to analyze and detect phishing on Twitter. The dataset included 11054 tweets. The approach comprised three steps, verification in a blacklist, the analysis of URLs, and the analysis of user accounts. Therefore, 25 URLs and 6 account features were extracted. Three ML algorithms were used, which are: LR, RF, and SVM. The results showed that the RF algorithm achieved the best results for both URL and account detection, with an accuracy of 95.51% and 75%, respectively.

In (Mughaid, A., 2022), the authors proposed an ML detection model to detect phishing in email. The dataset was split to train the detection model and validate the results using the test data. The work has been done in three phases using three different types of datasets, each set with different features. Since there are different features, different algorithms were applied and compared in terms of accuracy to find the most accurate and efficient results achieved. The results show that boosted DT algorithm has the highest accuracy of 88%, 100%, and 97% on the applied datasets. The authors stated that obtaining predetermined dataset is the only challenge for this study.

In (Wang, A.H., 2010), the authors proposed an ML spam detection model to detect spam bots on Twitter. The dataset was 500K tweets collected using different tools. Two types of features were used,

which are graph and content-based features. Four classification algorithms were tested, which are: DT, NN, SVM, KNN, and NB. Evaluation results show that the detection model is efficient in identifying spam bots on Twitter using the NB classifier, which achieved the best performance with 91.7% precision. This paper focused on twitter spam bots which could be unapplicable in other OSN platforms.

In (Ahmed, F., 2012), the authors presented an approach called Markov Clustering (MCL) to detect spam profiles on OSNs using a real dataset of Facebook containing spam and normal profiles. The social network was presented as a weighted graph to study the behavior of spammers, depending on different features such as posts, pages, and tags. Other features are active friends, page likes, and URLs. The main findings show that normal profiles have a maximum of 20-100 mutual friends. While spammers communicate with up to 600. Overall, the technique shows promise in detecting spam profiles but may require further exploration for spam campaign identification and comparison with supervised learning techniques.

In (Singh, M., 2016), the authors presented a behavioral analysis of spammers on Twitter using ML classification algorithms based on graph and content features. The dataset was about 74,000 tweets collected from 18,000 users. Five ML classification algorithms were utilized to classify spammers and genuine users. The experimental findings demonstrated that the RF classifier could detect spammers with a 91.96% accuracy (Thooyamani K.P., et.al, 2014). While the technique provides valuable insights, it is specific to Twitter and lacks exploration of other sorts of spammers. Furthermore, the report fails to address Twitter's policy flaws in recognizing pornographic users as spammers.

In (Alom, Z., 2018), the authors established a set of three novel graph-based and four content-based features discovered to detect spam accounts on Twitter. The researchers employed the following ML algorithms: KNN, DT, NB, RF, LR, SVM, and eXtreme Gradient Boosting (XGBoost). The dataset contained 41,499 accounts. 22,223 of the accounts were spammers, and 19,276 were legitimate users. The approach achieved an accuracy of 91% for the RF classifier and the lowest accuracy of 74% for the NB classifier. Also, the RF and XGBoost classifiers got the highest precision value of 92%. Despite the authors methodology hence using twitter make the generalization of the model for other OSN difficult.

Based on Dataset Size and Language

In this section, the reviewed studies are classified based on the Language corresponding to dataset size. Four ranges of the dataset size are used: Size< 500, 500<Size< 1000, 1000<Size< 5000, and Size>5000. Phishing detection based on the language of the dataset corresponding to the dataset size classified into:

- **English URLs and websites:** In [16, 21–26, 32], the authors used URLs and website datasets.
- **English dataset for detecting spam text or account:** In [14, 20, 29–31, 33, 41, 43–47, 52, 56–62] the authors used English account, text or text and account as a dataset for detecting English spam and English spam text.
- **Arabic dataset for detecting spam text or account:** In [34–38, 42, 49, 50, 53, 54], the authors used Arabic account, text, and account as the dataset for detecting Arabic spam and spam Arabic text.
- **Other datasets for detecting spam text or account:** In [27,51,48, 55], the authors used the Chinese account, or text and account as the dataset for detecting Chinese spam and spam Chinese text.

Table 1 summarizes the details of each dataset size for use in each of the reviewed articles according to the kind of language used In **Table 1**, we use 4 ranges of dataset size: size<500, 00<size<1000, 1000<size<5000, and size>50000. It is found that number of 8 articles aim to detect English URLs or

websites or URLs. It is found that the number of 23 articles aims to detect English text or accounts. On the other hand, 11 articles were intended for detecting Arabic spam content or spam accounts. Also, 3 articles were about Chinese spam content or accounts. It is shown that non-English dataset languages, such as Arabic and Chinese languages lack labeled datasets in the phishing field, for example, most research has ranged between 1000 to 5000 in Arabic dataset phishing detection, because of the difficulty of the labeling process for Arabic datasets. Additionally, due to its complicated morphology and syntax, the Arabic language can be a problem in machine learning. Because the Arabic language includes various dialects, it is hard to develop models that are accurate across all dialects. Finally, there are many phrases in the language that have several meanings, making it difficult for machines to correctly interpret the intended meaning. In (Mubarak, H., 2020), the results showed that the tool had difficulties in recognizing spam accounts when they post tweets that toggle between Arabic and English (Major Challenges of Natural Language Processing (NLP), 2023). **Figure 5** illustrates the distribution for reviewed articles based on data size corresponding to languages.

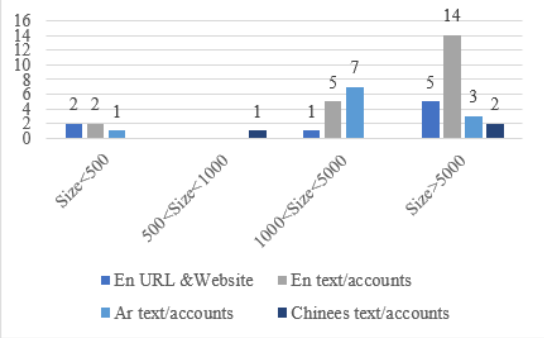

Figure 5: Number of Publications Based on Dataset Size and Corresponding Language

Table 1: Dataset Size According to the Language Type for Dataset

Language	Size< 500	500<Size< 1000	1000<Size< 5000	Size> 5000
English URLs and web Sites	[22, 26]	–	[23]	[16, 21, 24, 25, 32]
English text / Account	[59, 60]	–	[20, 43–46]	[14, 28–31, 33, 41, 47, 52, 56–58, 61, 62]
Arabic text / Account	[38]	–	[34–37, 42, 50, 54]	[40, 49, 53]
Others text / Account	–	[48]	–	[27,51, 55]

Based on Real-time or Offline Time

In this section, the reviewed studies are classified into two categories based on real-time prediction or not. Real-time prediction refers to making predictions or decisions in real-time as data is being generated, while offline prediction refers to making predictions or decisions on data that has already been collected. Real-time prediction is essential in applications where fast and immediate decisions are required, such as online fraud detection, real-time traffic monitoring, and weather forecasting. In this case, the system needs to process and analyze data in real-time to make accurate and timely predictions. On the other hand, offline prediction is suitable for applications where there is no urgency to make decisions, such as market analysis or customer segmentation. In such cases, data is collected and analyzed offline, and predictions are made based on historical data. Both real-time and offline prediction have their advantages and limitations, and the choice depends on the specific application and its requirements. **Table 2** summarizes the review articles according to the real−time, and offline time experiments results. Research [16, 21, 23, 25, 26, 41, 43, 46, 56] done their results in real time, whenever [14, 20, 24, 27–40, 42, 44, 45, 47–52, 54, 55, 57–62, 64, 65] the results were done in offline time.

Table 2: Dataset Size of Real-time Approaches vs Offline Time

Ref	Real time
[16, 21, 23, 25, 26, 41, 43, 46, 56]	**Yes**
[14, 20, 24, 27–40, 42, 44, 45, 47–52, 54, 55, 57–62, 64, 65]	No

Based on ML or DL

To prevent spam content, ML and DL approaches have been frequently deployed. However, in this context, they have major constraints and possible vulnerabilities. The capacity of ML and DL algorithms to recognize everchanging spam content is restricted. One of these limitations is the size of the dataset. It is difficult to have an appropriate amount of data for ML since a large dataset will result in a long training period and a small number will not yield correct results. On the other hand, DL approaches are notorious for overfitting, particularly when the dataset is short or unbalanced. Furthermore, to produce quicker results, both ML and DL need a considerable amount of computational power. **Figure 6** shows the number of studies conducted in either ML, DL or MCL at each category of URL based, content based, account based, and hybrid-based detection.

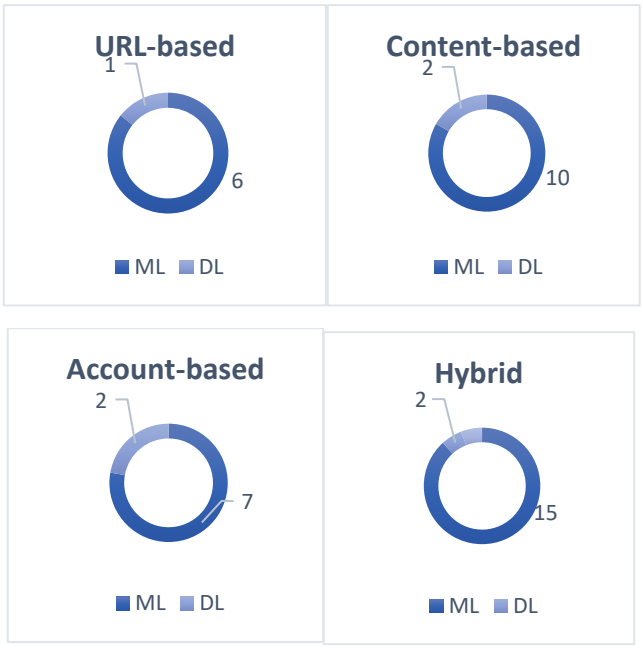

Figure 6: Research Articles using ML, DL

5 Discussion

A literature review of forty-five articles on OSN phishing detection using ML and DL of the period 2010 to 2022 is presented in **section 4**. The number of articles published each year is shown in **Figure 7**. After reviewing the related research based on the four different features (URL-based, content-based, and account-based features and hybrid-based features) there are some variations in the number of the related research for each feature type. **Figure 8** illustrates the percentage of the related research based on each feature. The highest rate is for hybrid features with 37.78% of total related research, while URL, content and account-based features take about 15.5%, 26.6%, and 20% of the total research consecutively. The majority of reviewed related research has examined phishing and spam in English language content on OSN extensively, while the Arabic language has received relatively less attention. As shown in **Figure**

8, the reviewed research based on URLs and websites is only 17.7%, Whereas Arabic spamming content research are only 24.4%, and the English language is 51% of the total number of related research. While there is 6.6% for the Chinese language. Also, the major difference for most of the related research the other is to implementation of the detection model in either a real-time tool or an offline tool. **Figure 8** shows that approximately 20% of reviewed articles detected phishing in real-time while approximately 80 % detected it in off-time. Tables 3, 4, 5, 6, and 7 summarize the related research based on the previous taxonomy, the features used for prediction, the dataset size corresponding language used, real-time based applications, and machine learning or deep learning techniques. Table 3 summaries of related research on spam detection based on URL based features (Nowakowski, P., 2021). Table 4 summaries of related research on spam detection based on content-based features. Table 5 summaries of related research on spam detection based on account-based features. Table 6, and Table 7 Summary of related research on spam detection based on hybrid features.

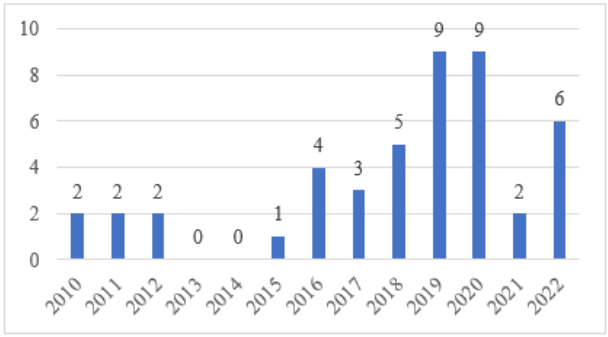

Figure 7: Number of Articles Published in each Year 2010-2022

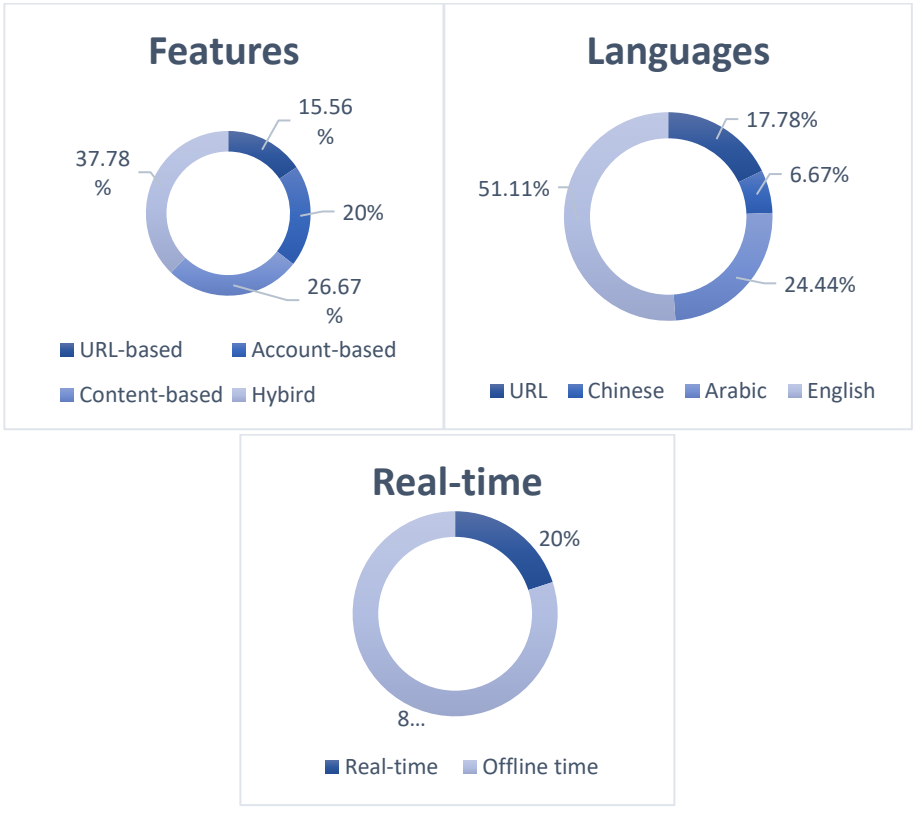

Figure 8: Research Articles Summary

Challenges and Opportunities

a) Challenges

All the mentioned literature reviews were about using machine learning in phishing detection. However, different challenges can face these techniques in real world. These challenges are listed as follows:

1. **Dataset size:** Data size is a significant challenge for ML-based detection techniques because the accuracy of ML models is affected. For example, for most reviewed articles in the Arabic language, the dataset size is small compared to the others. Where the dataset size for other articles is based on other languages, due to the labeling being tougher and needing manual labeling. If the dataset size is small, it may not contain enough data points to accurately represent the underlying patterns in the data. On the other hand, if the dataset is too large, it can take a long time to train and process the data, which can lead to slower training times and higher computational costs. As in (Alorini, D., 2019), the results showed that as the number of tested tweets increases, the accuracy decreases. In addition, the balancing of the dataset decreased the accuracy in (Mataoui, M.H., 2017).
2. **Real-time deployment:** This depends on the time complexity of the algorithm, which can affect the speed and accuracy of the algorithms used to train models. As the size of the data increases, the time complexity of algorithms can become a bottleneck, making it difficult to train models quickly **and** accurately. Therefore, the real-time deployment of machine learning models can be a challenge due to the time complexity of the models. In addition, it can be difficult due to the need for continuous monitoring and testing to ensure accuracy and reliability. As in (Sohrabi, M.K., 2018), the researchers had to use the lower accuracy algorithm to deploy the model in real time because it had a lower time complexity.
3. **Natural Language Processing (NLP):** For some languages, such as the Arabic language pose unique **challenges** due to their complex morphology, which is characterized by a rich system of inflections and derivations. For example, there are some of the major challenges and solutions in Arabic NLP such as morphological ambiguity, lack of standardized spelling, limited language resources, and bi-directional text.

b) Opportunities

There are several opportunities for predicting phishing attacks using ML and DL, such as the development of novel algorithms, real-world system development, and the use of multiple datasets from various sources.

1. **Development of hybrid features-based detection algorithms:** There aren't many algorithms that deal with the many kinds of data that are present in a dataset, comparable to the account features and text NLP, especially for the non-English text. Researchers can create new algorithms that function effectively on datasets with various forms of data to do more research.
2. **Increasing data size using data augmentation:** Data augmentation is a process of increasing the size and quality of a dataset by adding more data or modifying the data (Shorten, C., 2019). NLP data augmentation is the process of using various techniques and methods to increase the amount, variety, and quality of data available for training ML or DL models for NLP tasks. As the main challenge for more languages is the text dataset size, NLP data augmentation can be used for increasing the dataset to perform more accurately.
3. **Real-time deployment:** According to the studies evaluated in this paper, there hasn't been much progress in creating a practical method that can be followed consistently across various social media

platforms. To develop trustworthy automated systems that can foresee phishing attempts, more study in this area is urged.

4. **Addressing the challenges of language NLP:** Addressing the challenges of language NLP requires a combination of domain-specific knowledge, specialized tools, and the development of language resources. Continued efforts in building language resources and developing specialized NLP techniques will enable the development of more sophisticated applications and services that can support Arabic language users. Continued efforts in building language resources and developing specialized NLP techniques will enable the development of more sophisticated applications and services that can support prediction attack methods.

Table 3: Summary of Related Research on Spam Detection: URL Based Features

Ref/ Year	ML/DL	Classifier	Performance (Accuracy-F1)	Dataset	Real time
[16] (2019)	ML	DT, Adaboots, Kstar, KNN, RF, SMO, NB	DT based NLP: 97.02%, Adaboots based NLP: 93.24%, Kstar based hybrid: 95.27%, KNN (k=3) based hybrid: 95.86%, RF based NLP: 97.98%, 94. SMO based NLP: 94.92%, NB based hybrid: 95.86%	73,575 URLs	**Yes**
[21] (2011)	ML	BN	F1:0.9894	8118 phish and 4883 legitimate Web pages	**Yes**
[22] (2011)	ML	MLP, J48, NB	MLP:97%, J48:98.5%, NB:93.5%	200 URL and the Corresponding HTML	No
[23] (2015)	ML	NB, J48 DT, MLP, & BN	NB: 55%, J48 DT:97%, MLP: up to 97%, & BAYES NET: 66% (Accuracy within time after clicking the URL),	4000 English tweets contain URLs	**Yes**
[24] (2018)	ML	NB, SVM	NB:76.87% and SVM:91.28%	A set 23,000 phishing URLs and 2000 non-phishing URLs	No
[25] (2018)	DL	ANN, LSTM	ANN:95.57%, LSTM:96.89%	456300	**Yes**
[26] (2019)	ML	Random Forest (RF)	94.57%	200Phishing URLs	**Yes**

Table 4: Summary of Related Research on Spam Detection: Content Based Features

Ref/ Year	ML/DL	Classifier	Performance (Accuracy-F1)	Dataset	Real time
[27] (2022)	DL	A hybrid BiGRUCNN network	F1-score (over 94%)	Training 208,240, Dev 69,413, and Test 69,415 **Chinees** tweets	No
[28] (2020)	ML	Deep neural network (DNN)	F1: 70%	600 million **English** tweets, of which 6.5 million are malicious tweets	No
[29] (2017)	ML	KNN, Bagging, NB and SVM	KNN: 93%, Bagging: 94%, NB: 92% & SVM: 92 %	1,956 **English** Comments of YouTube Videos	No
[30] (2017)	DL	Binary ML classifier	92-99 %	600 million **English** tweets	No
[31] (2018)	ML	Clustering & SVM	Clustering: 70.8 & SVM: 92.5%	200,000 **English** Facebook posts	No
[32] (2021)	ML	SVM, NB, KNN, DT and RF	SVM: 98% , NB: 99% , KNN: 90% , DT: 94% & RF: 84%	5695 Emails	No
[33] (2022)	ML	Clustering Based on Peer acceptance	96.9 %	Three datasets:1. Social Honey Pot with size 2169, 2. HSpam14 with size 2000, 3. The Fake Project with size 1563 accounts	No
[34] (2022)	ML	Stacking Ensemble Classifier	96–99.5 %	1600 reviews translated from English to Arabic & 1600 **Arabic** reviews	No

[35] (2022)	ML	NB, SVM and LR	NB: 0.74%, LR: 0.74%, SVM: 0.923%	1648 **Arabic** spam tweets after augmentation.	No
[36] (2018)	ML	SVM, NB, & DT	87.32%, 82.42%, & 84.3%	1944 spam & 1559 legitimate **Arabic** tweets	No
[37] (2021)	ML	DT	92.63 %	3,000 **Arabic** Facebook comments	No
[38] (2020)	ML	Rule-based & NB	52% & 86%	150 WhatsApp & SMS **Arabic** messages	No

Table 5: Summary of Related Research on Spam Detection: Account Based Features

Ref/ Year	ML/DL	Classifier	Performance (Accuracy-F1)	Dataset	Real time
[39] 2022	DL	CNN text model combined with metadata model	Accuracy (94.27%)	1.2 million **Arabic** tweets for 16,700 users	No
[40] 2020	ML	New semi- supervised graph embedding model based on a graph. attention network	F1-score (91%)	Not Applied size, **English Twitter Account**	No
[41] (2010)	ML	DT, NN, SVM & NB	DT: 44.4%, NN: 58.8%, SVM: 40% and NB: 91.7%	25,847 Twitter users, around 500K **English** tweets, & around 49M follower/friend relationships	**Yes**
[42] (2019)	ML	LP & LS	91 %	1663 Twitter **Arabic** accounts	No
[43] (2019)	ML	LR with Clustering	99 %	1000 **English** Facebook users' data	**Yes**
[44] (2019)	DL	DT	90%	3000 Twitter **English** accounts	No
[45] (2019)	ML	DT & NB	DT: 99.28% & NB: 78.33%	2816 **English** Facebook accounts	No
[46] (2020)	ML	Similarity measures and DT	90 %	2980 Twitter **English** fake accounts	**Yes**
[47] (2021)	ML	Logistic Regression and Particle Swarm Optimization (PSO)	96%	The Fake Project DS: 6973 **English** Twitter accounts	No

Table 6: Summary of Related Research on Spam Detection: Hybrid Features_Part1

Ref/ Year	Features	ML/DL	Classifier	Performance (Accuracy-F1)	Dataset	Real time
[48] (2016)	Account & Content	ML	ELM, SVM, DT, NB, BN	ELM: 99.5%, SVM: 99.5%, DT: 94.7%, NB: 93% & BN: 92.6%	500 **Chinese** Sina Weibo Messages	No
[49] (2017)	Content & Account	ML	NB, DT, SMO, DT (T for Table), LR, & SGD	NB: 64.58%, DT: 76.57%, SMO: 68.31%, DT (T for Table): 72.44%, LR: 74.38%, & SGD: 69.39%	99 **Arabic** posts & 9697 Facebook comments	No
[20] (2019)	Account & Content	ML	SVM, DT, NB, KNN, & RF	SVM: 60%, DT: 68%, NB: 12%, KNN: 49%, & RF: 23%	1729 **English** Twitter accounts	No
[50] (2019)	Content	ML	NB & SVM	NB: 86% & SVM: 83%	2000 **Arabic** tweets	No
[51] (2019)	Account & Content	ML	SVM, DT, NN & NB	SVM: 98.1%, DT: 98.4%, NN: 0%, & NB: 98.9%	**Chinese** Sina Weibo Un-Specified size	No
[52] (2020)	Tweet & Profile	DL	MOEFS + RDNN	Hyves DS: 90.02%, Twitter DS: 95.60%	**English** 821 Hyves Messages & 61,675 tweets	No
[53] (2020)	Account & Content	ML	SVM & AraBERT	SVM: 99.5% & AraBERT: 99.7%	134,222 **Arabic** Tweets	No
[54] (2016)	Account & Content	ML	NB, RF, & SVM	NB: 87.24%, RF:92.59%, & SVM: 90.12%	5000 **Arabic** Tweets	No
[14] (2016)	Content & URL	ML	NB, SVM, & Integrated Approach	NB: 94%, SVM: 92%, & Integrated Approach: 98%	15000 URL & 100 **English** users' tweets	No

Table 7: Summary of Related Research on Spam Detection: Hybrid Features_Part2

Ref/ Year	Features	ML/DL	Classifier	Performance (Accuracy-F1)	Dataset	Real time
[55] (2012)	Forwarding, URL, & Account	ML	NB, DT, RF	BN: 84.74%, DT: 82.22%, RF: 79.07%	12,006 **Chinese** Sina Weibo message	No
[56] (2020)	URL, WHO-Is, tweet, & network	ML	NB, DT, RF	NB: 87%, DT: 89.2%, & RF: 92.5%	309,321 **English** tweets	**Yes**
[57] (2020)	URL & Account	ML	LR, SVM, RF	LR: 90.28%, SVM: 93.43%, & RF: 95.51%	11,054 **English** account & URL	No
[58] (2022)	Different Features	ML	Locally deep SVM, SVM, Boosted DT, LR, Averaged perceptron, NN, DF	Locally deep SVM: 99.5%, SVM: 99.7%, Boosted. DT: 100%, LR: 99.8%, Averaged perceptron: 99.6%, NN: 99.5%, DF: 99.9%	10,000 **English** Emails	No
[59] (2010)	Account & Content	ML	DT, NN, SVM, NB	DT: 44.4%, NN: 58.8%, SVM: 40%, NB: 91.7%	500 **English** Accounts	No
[60] (2012)	Account	DL	Markov Clustering (MCL)	FP: 88% & FB: 79%	320 Facebook **English** Accounts	No
[61] (2016)	Graph & Content	ML	NB, LR, DT, RF, AdaBoostM1	NB: 84.5%, LR: 82.3%, DT: 89.2%, RF: 91.9%, AdaBoostM1: 83.8%.	74,000 **English** Tweets	No
[62] (2018)	Account	ML	KNN, DT, NB, RF, LR, SVM, and XG-Boost	KNN: 92%, DT: 91%, NB: 74%, RF: 92%, LR: 89%, SVM: 83%, XGBoost: 91%	41,499 **English** Facebook, Twitter & MySpace Accounts	No

6 Conclusion

This paper reviewed 45 studies related to phishing detection over OSNs. After gathering articles, they divided into three main groups being able to make the review more understandable: features-based prediction, real-time deployment system prediction and dataset language utilized in the prediction. After that, each category had investigated. A detailed comparison of the reviewed studies had illustrated in tables. We found that most of the studies that have been proposed focused on developing predictive models aimed at predicting phishing attacks based on hybrid features such as account and text spam detection for the English test dataset compared with another language. Finally, we concluded by outlining challenges associated with exploring issues and opportunities in the realm of ML and DL applications in phishing attack prediction. Future researchers should be considerate about spam detection for non-English languages by developing specialized NLP techniques to generate more sophisticated applications and services that can support prediction attack methods. Also, more study for OSNs real time spam detection systems should be in developed, to foresee phishing attempts in real times.

References

[1] Aggarwal, A., Rajadesingan, A., & Kumaraguru, P. (2012). PhishAri: Automatic realtime phishing detection on twitter. *In IEEE ECrime researchers summit*, 1-12.

[2] Ahmed, F., & Abulaish, M. (2012). An mcl-based approach for spam profile detection in online social networks. *In IEEE 11th international conference on trust, security and privacy in computing and communications*, 602-608.

[3] Al-Azani, S., & El-Alfy, E.S.M. (2018). Detection of arabic spam tweets using word embedding and machine learning. *In IEEE international conference on innovation and intelligence for informatics, computing, and technologies (3ICT)*, 1-5.

[4] Alharbi, A., Alotaibi, A., Alghofaili, L., Alsalamah, M., Alwasil, N., & Elkhediri, S. (2022). Security in social-media: Awareness of Phishing attacks techniques and countermeasures. *In IEEE 2nd International Conference on Computing and Information Technology*, 10-16.

[5] Alharthi, R., Alhothali, A., & Moria, K. (2019). Detecting and characterizing arab spammers campaigns in twitter. *Procedia Computer Science*, *163*, 248-256.

[6] Alhassun, A.S., & Rassam, M.A. (2022). A combined text-based and metadata-based deep-learning framework for the detection of spam accounts on the social media platform twitter. *Processes*, *10*(3), 1-24.

[7] Alkadri, A.M., Elkorany, A., & Ahmed, C. (2022). Enhancing detection of arabic social spam using data augmentation and machine learning. *Applied Sciences*, *12*(22), 1-16.

[8] Alom, Z., Carminati, B., & Ferrari, E. (2018). Detecting spam accounts on Twitter. *In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1191-1198.

[9] Alorini, D., & Rawat, D.B. (2019). Automatic spam detection on gulf dialectical Arabic Tweets. *In IEEE international conference on computing, networking and communications (ICNC)*, 448-452.

[10] Baatarjav, E.A., Phithakkitnukoon, S., & Dantu, R. (2008). Group recommendation system for facebook. *In On the Move to Meaningful Internet Systems: OTM 2008 Workshops: OTM Confederated International Workshops and Posters, ADI, AWeSoMe, COMBEK, EI2N, IWSSA, MONET, On To Content+ QSI, ORM, PerSys, RDDS, SEMELS, and SWWS 2008, Monterrey, Mexico. Proceedings*, 211-219. Springer Berlin Heidelberg.

[11] Barbier, G., & Liu, H. (2011). Data mining in social media. *Social network data analytics*, 327-352.

[12] Barushka, A., & Hajek, P. (2020). Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks. *Neural Computing and Applications*, *32*, 4239-4257.

[13] Bharti, K.K., & Pandey, S. (2021). Fake account detection in twitter using logistic regression with particle swarm optimization. *Soft Computing*, *25*(16), 11333-11345.

[14] Boreggah, B., Alrazooq, A., Al-Razgan, M., & AlShabib, H. (2018). Analysis of arabic bot behaviors. *In IEEE 21st Saudi Computer Society National Computer Conference (NCC)*, 1-6.

[15] Burnap, P., Javed, A., Rana, O.F., & Awan, M.S. (2015). Real-time classification of malicious URLs on Twitter using machine activity data. *In Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 970-977.

[16] Cao, J., Li, Q., Ji, Y., He, Y., & Guo, D. (2016). Detection of forwarding-based malicious URLs in online social networks. *International Journal of Parallel Programming*, *44*, 163-180.

[17] Chy, M.K.A., Ahmed, S.A., Doha, A.H., Masum, A.K.M., & Khan, S.I. (2019). Social media user's safety level detection through classification via clustering approach. *In IEEE International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 1-4.

[18] Cortizo, J., Carrero, F., Gomez, J., Monsalve, B., & Puertas, E. (2009). Introduction to Mining SM. In *Proceedings of the 1st International Workshop on Mining SM*, 1-3.

[19] Delle, F.A., Clayton, R.B., Jordan Jackson, F.F., & Lee, J. (2022). Facebook, Twitter, and Instagram: Simultaneously examining the association between three social networking sites and relationship stress and satisfaction. *Psychology of Popular Media*.

[20] Djaballah, K.A., Boukhalfa, K., Ghalem, Z., & Boukerma, O. (2020). A new approach for the detection and analysis of phishing in social networks: the case of Twitter. *In IEEE Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 1-8.

[21] El-Mawass, N., & Alaboodi, S. (2016). Detecting Arabic spammers and content polluters on Twitter. *In IEEE Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*, 53-58.

[22] Elyusufi, Y., Elyusufi, Z., & Kbir, M.H.A. (2019). Social networks fake profiles detection based on account setting and activity. *In Proceedings of the 4th International Conference on Smart City Applications*, 1-5.
[23] Gilwit, D.B. (2003). The Latest Cybersquatting Trend: Typosquatters, Their Changing Tactics, and How to Prevent Public Deception and Trademark Infringement. *Wash. UJL & Pol'y*, *11*.
[24] Goel, D., & Jain, A.K. (2018). Mobile phishing attacks and defence mechanisms: State of art and open research challenges. *computers & security*, *73*, 519-544.
[25] Ho, K., Liesaputra, V., Yongchareon, S., & Mohaghegh, M. (2018). Evaluating social spammer detection systems. *In Proceedings of the Australasian Computer Science Week Multiconference*, 1-7.
[26] Jain, A.K., & Gupta, B.B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. *In Cyber Security: Proceedings of CSI*, 467-474. Springer Singapore.
[27] Jose, T., & Babu, S.S. (2019). Detecting spammers on social network through clustering technique. *Journal of Ambient Intelligence and Humanized Computing*, 1-15.
[28] Kandasamy, K., & Koroth, P. (2014). An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques. *In IEEE Students' Conference on Electrical, Electronics and Computer Science*, 1-5.
[29] Kao, A., & Poteet, S.R. (Eds.). (2007). *Natural language processing and text mining*. Springer Science & Business Media.
[30] King, I., Li, J., & Chan, K. T. (2009). A brief survey of computational approaches in social computing. *In IEEE International Joint Conference on Neural Networks*, 1625-1632.
[31] Koggalahewa, D., Xu, Y., & Foo, E. (2022). An unsupervised method for social network spammer detection based on user information interests. *Journal of Big Data*, *9*(1), 1-35.
[32] Kontsewaya, Y., Antonov, E., & Artamonov, A. (2021). Evaluating the effectiveness of machine learning methods for spam detection. *Procedia Computer Science*, *190*, 479-486.
[33] Lakshmi, V.S., & Vijaya, MS. (2012). Efficient prediction of phishing websites using supervised learning algorithms. *Procedia Engineering*, *30*, 798-805.
[34] Liew, S.W., Sani, N.F.M., Abdullah, M.T., Yaakob, R., & Sharum, M.Y. (2019). An effective security alert mechanism for real-time phishing tweet detection on Twitter. *Computers & security*, *83*, 201-207.
[35] Major Challenges of Natural Language Processing (NLP), (2023). https://monkeylearn.com/blog/natural-language-processing-challenges/.
[36] Marie-Sainte, S.L., Alalyani, N., Alotaibi, S., Ghouzali, S., & Abunadi, I. (2018). Arabic natural language processing and machine learning-based systems. *IEEE Access*, *7*, 7011-7020.
[37] Mataoui, M.H., Zelmati, O., Boughaci, D., Chaouche, M., & Lagoug, F. (2017). A proposed spam detection approach for Arabic social networks content. *In IEEE International Conference on Mathematics and Information Technology (ICMIT)*, 222-226.
[38] Mourad, A., Srour, A., Harmanani, H., Jenainati, C., & Arafeh, M. (2020). Critical impact of social networks infodemic on defeating coronavirus Covid-19 pandemic: Twitter-based study and research directions. *IEEE Transactions on Network and Service Management*, *17*(4), 2145-2155.
[39] Mubarak, H., Abdelali, A., Hassan, S., & Darwish, K. (2020). Spam detection on arabic twitter. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, Proceedings 12*, 237-251. Springer International Publishing.
[40] Mughaid, A., AlZu'bi, S., Hnaif, A., Taamneh, S., Alnajjar, A., & Elsoud, E.A. (2022). An intelligent cyber security phishing detection system using deep learning techniques. *Cluster Computing*, *25*(6), 3819-3828.
[41] Najadat, H., Alzubaidi, M.A., & Qarqaz, I. (2021). Detecting Arabic spam reviews in social networks based on classification algorithms. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1-13.

[42] Najadat, H., Alzubaidi, M.A., & Qarqaz, I. (2021). Detecting Arabic spam reviews in social networks based on classification algorithms. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1-13.

[43] Nowakowski, P., Zórawski, P., Cabaj, K., & Mazurczyk, W. (2021). Detecting Network Covert Channels using Machine Learning, Data Mining and Hierarchical Organisation of Frequent Sets. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 12*(1), 20-43.

[44] Pais, S., Cordeiro, J., & Jamil, M.L. (2022). NLP-based platform as a service: a brief review. *Journal of Big Data*, *9*(1), 1-26.

[45] Rahman, M.S., Halder, S., Uddin, M.A., & Acharjee, U.K. (2021). An efficient hybrid system for anomaly detection in social networks. *Cybersecurity*, *4*(1), 1-11.

[46] Razzaq, A., Hur, A., Ahmad, H.F., & Masood, M. (2013). Cyber security: Threats, reasons, challenges, methodologies and state of the art solutions for industrial applications. *In IEEE Eleventh International Symposium on Autonomous Decentralized Systems (ISADS)*, 1-6.

[47] Saeed, R.M., Rady, S., & Gharib, T.F. (2022). An ensemble approach for spam detection in Arabic opinion texts. *Journal of King Saud University-Computer and Information Sciences*, *34*(1), 1407-1416.

[48] Sahingoz, O.K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345-357.

[49] Seifert, J. W. (2004). Data mining: An overview. *National security issues*, 201-217.

[50] Sharmin, S., & Zaman, Z. (2017). Spam detection in social media employing machine learning tool for text mining. *In IEEE 13th international conference on signal-image technology & internet-based systems (SITIS)*, 137-142.

[51] Shivangi, S., Debnath, P., Sajeevan, K., & Annapurna, D. (2018). Chrome extension for malicious URLs detection in social media applications using artificial neural networks and long short-term memory networks. *In IEEE international conference on advances in computing, communications and informatics (ICACCI)*, 1993-1997.

[52] Shorten, C., & Khoshgoftaar, T.M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1-48.

[53] Singh, M., Bansal, D., & Sofat, S. (2016). Behavioral analysis and classification of spammers distributing pornographic content in social media. *Social Network Analysis and Mining*, *6*, 1-18**.**

[54] Sohrabi, M.K., & Karimi, F. (2018). A feature selection approach to detect spam in the Facebook social network. *Arabian Journal for Science and Engineering*, *43*(2), 949-958.

[55] Sowmya, P., & Chatterjee, M. (2020). Detection of fake and clone accounts in twitter using classification and distance measure algorithms. *In IEEE International Conference on Communication and Signal Processing (ICCSP)*, 0067-0070.

[56] Tandon, A., Guha, S.K., Rashid, J., Kim, J., Gahlan, M., Shabaz, M., & Anjum, N. (2022). Graph based CNN algorithm to detect spammer activity over social media. *IETE Journal of Research*, 1-11.

[57] Thooyamani K.P., et.al (2014). Deploying site-to-site VPN connectivity: MPLS Vs IPSec. *World Applied Sciences Journal, 29*(14), 6-10.

[58] Wang, A.H. (2010). Detecting spam bots in online social networking sites: a machine learning approach. *In IFIP Annual Conference on Data and Applications Security and Privacy*, 335-342. Berlin, Heidelberg: Springer Berlin Heidelberg.

[59] Wang, A.H. (2010). Don't follow me: Spam detection in twitter. *In IEEE international conference on security and cryptography (SECRYPT)*, 1-10.

[60] Wang, B., Zubiaga, A., Liakata, M., & Procter, R. (2015). Making the most of tweet-inherent features for social spam detection on Twitter.

[61] Wei, F., & Nguyen, U.T. (2019). Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. *In First IEEE International conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)*, 101-109.

[62] Wu, T., Wang, D., Wen, S., & Xiang, Y. (2017). How spam features change in Twitter and the impact to machine learning based detection. In *Information Security Practice and Experience: 13th International Conference, ISPEC 2017, Melbourne, VIC, Australia, Proceedings 13*, 898-904. Springer International Publishing.

[63] Xiang, G., Hong, J., Rose, C.P., & Cranor, L. (2011). Cantina+ a feature-rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TISSEC)*, *14*(2), 1-28.

[64] Yao, J., Wang, C., Hu, C., & Huang, X. (2022). Chinese spam detection using a hybrid BiGRU-CNN network with joint textual and phonetic embedding. *Electronics*, *11*(15), 1-15.

[65] Zhao, C., Xin, Y., Li, X., Yang, Y., & Chen, Y. (2020). A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data. *Applied Sciences*, *10*(3), 1-18.

[66] Zhao, C., Xin, Y., Li, X., Zhu, H., Yang, Y., & Chen, Y. (2020). An attention-based graph neural network for spam bot detection in social networks. *Applied Sciences*, *10*(22), 1-15.

[67] Zheng, X., Zhang, X., Yu, Y., Kechadi, T., & Rong, C. (2016). ELM-based spammer detection in social networks. *The Journal of Supercomputing*, *72*, 2991-3005.

Authors Biography

Azza A. Abdo was born in Egypt, in 1982. She received the B.S., M.Sc., and Ph.D. degrees in computer science from the Faculty of Science, Menoufia University, Egypt, in 2003, 2008, and 2012, respectively. She is currently an Assistant Professor with the Department of Mathematics and Computer Science, Faculty of Science, Menoufia University. She is also a Visitor with the Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Saudi Arabia since 2018. Here research interests include the areas of cryptography, network security, application of chaotic systems in multimedia content encryption, Signatures, and authentications. Email: aaaali@iau.edu.sa , and https://orcid.org/0000-0001-8246-9790

Khaznah Alhajri is undergraduate student at computer science. Imam Abdulrahman Bin Faisal University (IAU), Saudi Arabia. Actively engage in academic projects and contributions. Eager to contribute to the computer field, as she published various scientific papers in the field of computer science. Her research interests include the areas of networks, security, data science and applications development. The author can be reached at this e-mail address: 2190004752@iau.edu.sa , and the ORCID profile can be found at: https://orcid.org/0000-0003-0987-8058

Assail Alyami birthplace is Saudi Arabia, and she was born in the year 2000, is undergraduate student at College of Science and Humanities, Imam Abdulrahman Bin Faisal University (IAU) in computer science major. An enthusiastic learner with an insatiable thirst for knowledge in the realm of computer science. Yielded many projects under diverse topics during study such as multilingual programming, networking, data structure, and algorithms. Email: 2190005693@iau.edu.sa , and https://orcid.org/0009-0005-7416-2293

Aljazi Alkhalaf is undergraduate student at Imam Abdulrahman Bin Faisal University (IAU, Saudi Arabia. Holding a bachelor's degree in computer science with honours since 2023. Passionate learner of computer science fields, concepts, and approaches. Governance risk and compliance, cybersecurity, and data analysis are all her research interests. She acquired valuable knowledge by publishing scientific papers, training security analysts, and developing mobile apps. The author can be reached at this e-mail address: 2190001702@iau.edu.sa , and the ORCID profile can be found at: https://orcid.org/0000-0003-2586-836X .

Bashayer Allail is undergraduate student at Computer Science from Imam Abdulrahman Bin Faisal University (IAU), Saudi Arabia. She developed an early interest in computers and technology and pursued her passion by obtaining a solid foundation in programming, data structures, algorithms, and various other computer science concepts during her time at university. Emails: 2190004502@iau.edu.sa , and https://orcid.org/0000-0002-5693-6356

Esra Alyami, born in 2000, she is undergraduate student at Computer Science from Imam Abdulrahman Bin Faisal University (IAU). I worked on various projects covering computer science concepts, gaining practical experience in software development and other areas. For further information: ORCID profile: at https://orcid.org/0000-0003-0395-1605. Email: 2190005556@iau.edu.sa

Hind baageel was born in Saudi Arabia, in 1994. She got her bachelor's degree in computer science from jubail university college. She has a master's degree in information security from Imam Abdullrahman Bin Faisal university. For the past 3 years, she has been working as a lecturer in computer department in Imam Abdullrahman bin faisal university. Her main research interests are security in biometric systems, Application of AI in authentication and information security and ethical aspects of security. E-mail: haaalssayed@iau.edu.sa,
Orcid: https://orcid.org/0000-0002-6343-6114