Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

DATA-DRIVEN CONTENT-BASED WEB SPAM DETECTION: A SYSTEMATIC REVIEW OF DATA SCIENCE AND MACHINE LEARNING APPROACHES

Asim Shahzad[*1], Muhammad Shehzad Khan[2], Raees ur Rehman[3], Muhammad Naeem[4], Zobia Shabeer[5]

[*1,2,4,5]*Department of Computer Science, Abbottabad University of Science and Technology, Abbottabad, Pakistan*
[3]*The University of Lahore, Pakistan*

[*1]asim.shaz@gmail.com, [*2]me_shehzadkhan@yahoo.com, [*3]raeesuol@gmail.com [*4]naeem@aust.edu.pk, [5]szubia033@gmail.com

DOI: https://doi.org/10.5281/zenodo.21720165

Keywords
Data science; content-based web spam detection; spamdexing; machine learning; feature engineering; web-page classification; systematic literature review

Article History
Received: 02 November 2025
Accepted: 16 December 2025
Published: 30 December 2025

Corresponding Author: *
Muhammad Shehzad Khan

Abstract
Web spam uses deceptive or misleading techniques that are misleading in order to change how a website ranks on the search engine result pages (SERPs) and make finding useful information more difficult. This paper provides a systematic literature review on the subject of web spam recognition via a data science perspective. The paper provides an insight into the manner in which the data of a website is represented through text, HTML, semantic information, website structure, and link-related signals. The discussion concerning the techniques of data processing, feature engineering, data classification, machine learning, ensemble learning, and deep learning is also carried on. The reviewed studies show that data-driven performance depends on the choice of datasets, feature spaces, learning algorithms, class balance, and evaluation measures such as accuracy, precision, recall, and F-measure. The article also summarizes the strengths and weaknesses of existing content-based and combined detection approaches and identifies continuing challenges, including evolving spam strategies, limited benchmark diversity, computational cost, model generalization, and adversarial behavior. By organizing the literature around data, features, models, and evaluation, the review positions content-spam detection as an applied data science problem and highlights directions for more adaptive and reliable web spam detection.

I. INTRODUCTION

Web spam refers to a type of marketing over the internet whereby dishonest methods are employed in order to rank a website on the search engines. This can include designing or manipulating web pages and a host of other methods like paid inclusion and link farms. The ways in which web spam can lead to the degradation of search results quality are enormous. First, it could make it harder for people to find reliable information. When web spam is present in the SERPs, it can push down high-quality pages and make them more difficult to find. Second, web spam can waste users' time by leading them to irrelevant or low-quality pages. Third, web spam can damage the reputation of search engines by making users question the accuracy of their results [1].

Getting a website to show up high in search results means more people will visit, and that can lead to more revenue. But some people try to cheat the system with web spam. It's a fake and dishonest way to trick search engines into ranking their site better. Web spam ruins the quality of search results, wastes people's time, and makes them lose faith in the search engine. Because of this, researchers have done a lot of work to find and fight web spam[2].

Content-based web-spam detection analyzes the content of web pages for spammy or malicious elements, such as suspicious links, deceptive language, or irrelevant content. Behavioral-based spam detection instead watches how people interact with those pages over time, and then it flags weird or suspicious patterns, for example, users clicking links that end up on harmful websites, or entering personal information into websites that are not trusted. Spam detection is essential for protecting websites and ensuring that only legitimate content is delivered to intended visitors.

Content-based spam detection is one way to detect and remove unsolicited web pages, which are generated in large quantities and sent to many websites. It entails examining the web page content to look for spammy or virus-like aspects, such as suspicious links, language, and so on. It is an important technique for organizations and individuals who want to prevent spam attacks on their websites and have their genuine content reach their legitimate customers. In this technique, the spam web pages are identified and classified accurately using advanced machine learning algorithms, NLP techniques and other computational methods. In summary, content-based spam detection is a crucial and effective method that contributes to a more secure and reliable web environment, fostering users' trust and confidence. Mukunthan [3] studied content-based features for spam detection, like hashtags (#), URLs (#), @ mentions, retweets, spam terms, HTTP links, hot topics, and duplicate tweets, all of which have a part in how many people use Twitter. The factors to be considered in spam classification using a content-based approach are Keyword Stuffing, Doorway Pages, Body Spamming, Scraper Sites, Article Spinning, Title Spamming, URL Spamming, and Anchor Text Spamming [4]. The task can be approached from a data science point of view as a classification pipeline where the data from websites are preprocessed, transformed into meaningful features and assessed using an appropriate learning algorithm and performance measure. This workflow is summarized in Figure 1.

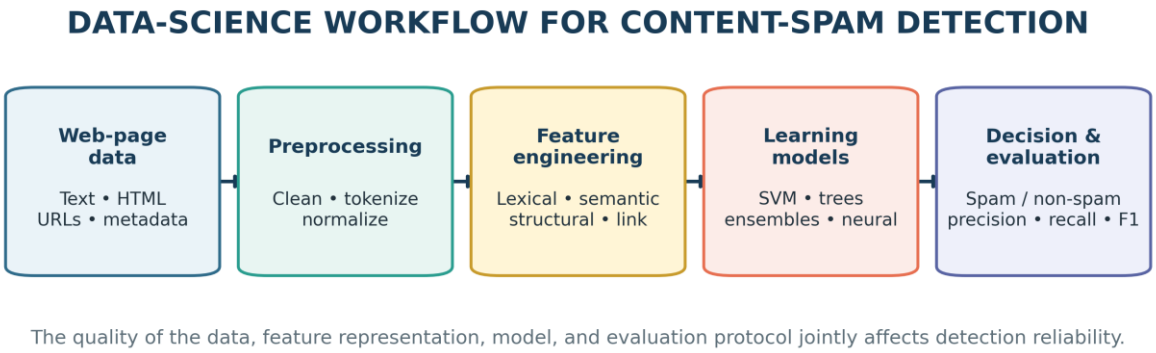

Figure 1. Data-science workflow for content-based web spam detection.

We organized this article as follows. Section 2 gives an overview of search engine optimization and its types. Section 3 introduces content-based web spam and summarizes the main categories. Then Section 4 provides a more detailed look at data-driven content-based and combined web spam detection algorithms. Section 5 presents a synthesis of the strengths and weaknesses in the

reviewed literature. After that, Section 6 discusses the continuing problems in identifying spam, followed by the conclusion in Section 7.

II. SEARCH ENGINE OPTIMIZATION

SEO (Search Engine Optimization) is basically a group of practices that help a website show up higher on search engine results pages, or SERPs, so it becomes more visible, and users are more likely to click. SEO also helps to ensure that users are presented with the most relevant and accurate information possible [5]. In many conversations, the term SEO can turn into a misnomer, since it is really a combination of on-page optimization and off-page optimization. As you can see in Figure 2, there's a correlation between on-page and off-page optimization; there's a line where manipulative actions begin to be considered content spam or link spam.

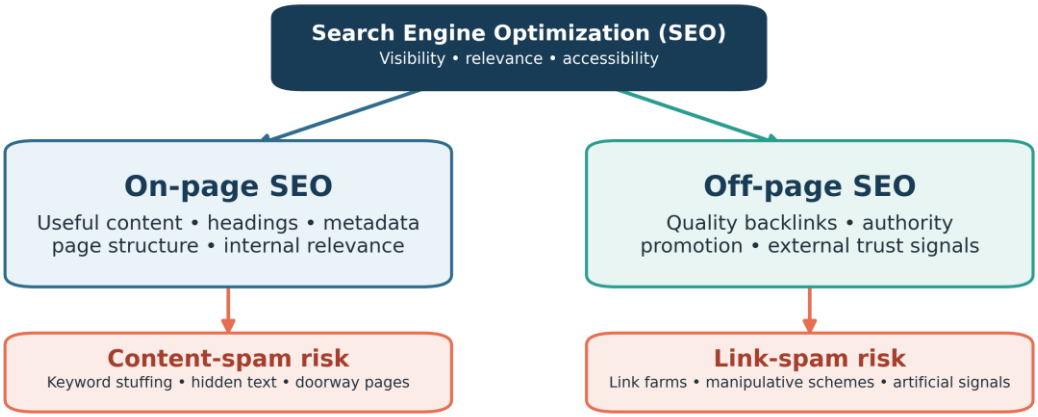

Figure 2. On-page and off-page SEO in relation to content-spam and link-spam risks.

On-page optimization focuses on things like high-quality content, carefully chosen keywords, placing those keywords in relevant tags, and then monitoring keyword frequency. For off-page optimization, white hat SEO is a legit and ethical approach, one that respects search engine guidelines, and it aims to increase both visibility and accessibility of a site in the search results pages (SERPs) [4]. The article [6] outlines unethical SEO practices, usually called black hat SEO, which are used to trick search engines so websites rank higher than they should. Typical examples include keyword stuffing, doorway pages, copying content, hidden text or links, and cloaking. Search engines keep improving their algorithms to spot and punish black hat SEO.

In [7] The authors did a careful review of the SEO literature on ranking factors, and picked out 24 key website characteristics, including on-page and off-page factors. Looking specifically at on-page factors, they found that having keywords in the title tag and keyword density in different parts of the web page were the most important factors. In [8] the authors used expert opinions, gathered via interviews and surveys, to highlight what they considered the most important SEO factors. Meta tags were described as the top factor, followed by keywords in the content, and then website design.

III. CONTENT-BASED WEB SPAM

Content spamming is a form of manipulating what's on a web page to improve placement in search engine results pages (SERPs), often done by relying on hyperlinks, high word count, and misleadingly repeating content [9]. There are various signals that can help spot this, such as stop words, keyword density, a spam keywords database, the ratio of parts of speech (POS), and even copied material, which can all be used for detection [10]. An article [11] looked at how to sort web pages into fixed, predefined categories

by using machine learning approaches, plus features taken from both the text and the HTML tags. The outcomes suggest that the count of HTML tags is important to identifying the page topic, but those tags aren't used as broadly in classification methods as they really could be.

Content spam itself can be grouped in different ways, mostly depending on what factors are being emphasized. In this review, the main content-spam categories are shown in Figure 3.

- Keyword Stuffing: Putting keywords in a "planned" way all over a web page so the keyword count, wording variation, and density go up is called keyword stuffing [10]. If a page uses too many keywords, or a bunch of close synonyms, then the web page becomes spammy, and that can even drag the ranking down in SERPs [12].
- Doorway Pages: These are low-quality pages with very little content that are designed to redirect users to another website [5].
- Body Spamming: This is a common style of content spam that changes the page body. It's easy and costs less, and it gives plenty of room for different tactics. For instance, a spammer may attempt to make a page rank highly for a specific search query by repeatedly using those queries several times across the main body text [13].
- URL spamming: This is a black hat SEO practice where the URLs are created in a way intended to tweak search engine rankings. For instance spammer constructs a URL that repeats one keyword multiple times, or that includes keywords that don't really match the page content at all [13].
- Meta-tag spam: Also considered a black hat SEO trick, it uses the idea of stuffing meta tags with keywords as a way to meddle with search engine rankings [13].

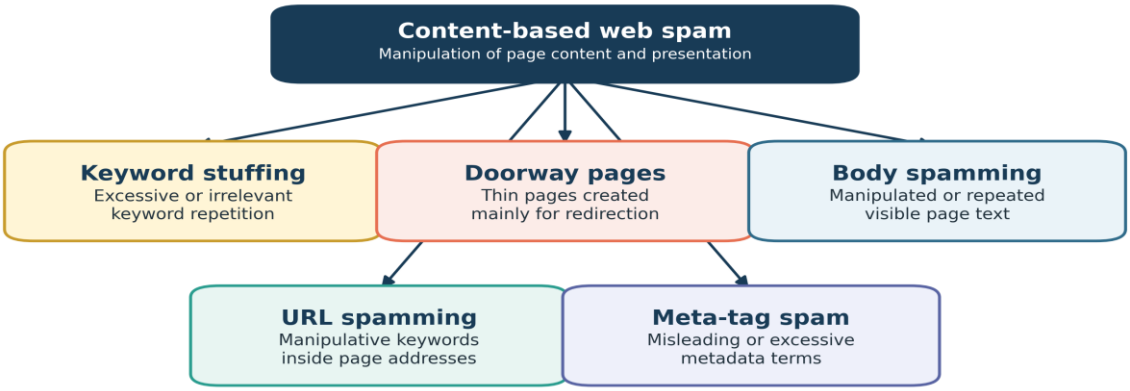

Figure 3. Taxonomy of content-based web spam considered in the review.

IV. DATA DRIVEN AND MACHINE LEARNING TECHNIQUES FOR COMBATING CONTENT AND COMBINED WEB SPAMMING

The reviewed researchers provide different data-driven approaches for combating content-based web spamming. These approaches can be organized by their data sources, feature representations, learning models, and validation methods, as summarized in Figure 4.

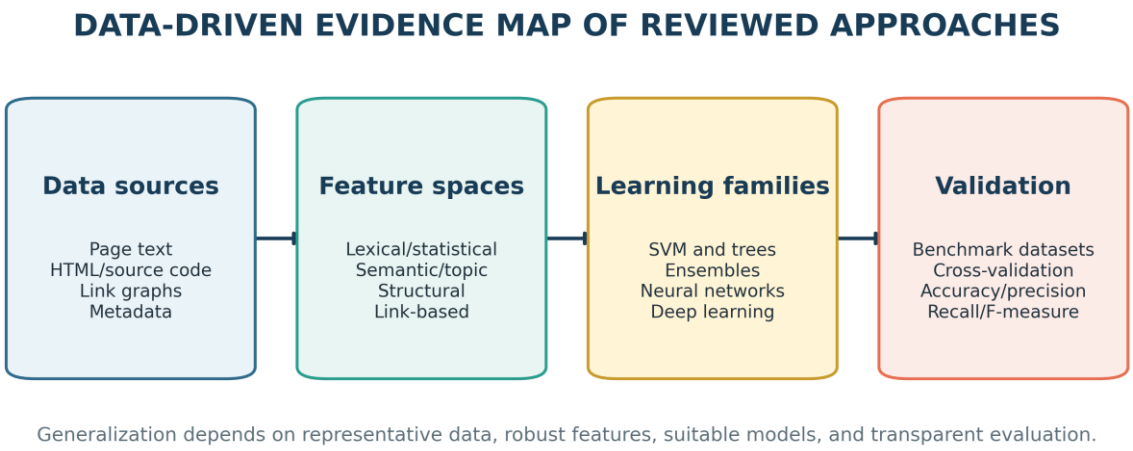

Figure 4. Data-driven evidence map of the web spam detection approaches reviewed in this study.

Zhuang et al. [14] introduced a new method for demoting web-spam pages using deep belief networks (DBNs). DBNs are a kind of artificial neural network that can learn to spot intricate patterns in data, in a way that feels almost like a multi-step recognition process. A DBN is trained with a bundle of web pages, both non-spam and spam. After training, the DBN is then used to estimate how likely a fresh web page is to be spam, basically a probability-style score. Pages that come out as high risk get demoted inside the SERPs. Two data collections, namely WEBSPAM-UK2006 and WEBSPAM-UK2007, have been used for evaluating the proposed approach. The results showed that the approach improved SERP quality by pushing down web spam sites in the rankings. They achieved 0.7% relative advancement on WEBSPAM-UK2006 and 7.3% relative enhancement on the WEBSPAM-UK2007 dataset, and reported an accuracy of 0.94, a precision of 0.95, and a recall of 0.95. Hua et al. [15] presented a novel scheme to detect spam based on correlations among content features of web pages. The researchers observed that current spam filtering algorithms do not perform well enough, as they are generally based on a single feature or only on a few features that can be circumvented by spammers. The proposed approach extracts a set of features from a web page such as title, keywords, description, body text, and links. The authors then performed a correlation analysis to determine correlations among these features. They claim that there are more correlations between features in spam pages than in legitimate pages, and these can be used to distinguish spam pages from legitimate pages. The authors tested their method on 10,000 web pages, 50% legitimate and 50% spam. The proposed method's accuracy of spam page detection was 95%. This is much more accurate than any other spam detection technology currently available. Asdaghi et al. [16] proposed a new backward elimination method for feature selection in web spam detection. The main idea is to measure the impact of removing a group of features on the performance of a classifier, instead of judging the effect of deleting just one feature at a time. The proposed approach was tested on the WEBSPAM-UK2007 dataset using Naïve Bayes classifiers and demonstrated to select a smaller number of features and to enhance the classifier's performance.

Liu et al. [17] came up with a novel approach to detect web spam using features of the home page source code. The authors claim that the current techniques for detecting web spam are insufficient, since they simply concentrate on the content and links found in websites. Spammers can, however, easily get past these detection methods, such as keyword stuffing and link farming. The proposed approach had an accuracy of 97.5%, a precision of 98.0%, and a recall of

97.0% on the given dataset of labeled web pages. Makkar et al. [18] presented a fresh web spam detection approach in an IoT setting using a deep learning technique. They used a recurrent neural network, specifically the Long Short-Term Memory (LSTM) network, and it is quite effective when working with ordered or time-based sequences, including text. The proposed scheme was tested on a dataset of 111k web pages, where 55,500 web pages were spam and 55,500 web pages were legitimate. The accuracy rate for identifying spam web pages was 95.25% in the scheme.

The [19] suggested the use of graph mining and machine learning to detect web spam. The authors found that web spam could be identified by analyzing the graph properties of the websites, including the number of incoming and outgoing links, the distribution of link weights, and the existence of a specific pattern of links. To test their approach, the authors used the WEBSPAM-UK2007 dataset. From the dataset, they extracted various graph features and built a classifier using five graph machine learning algorithms (SVM, KNN, logistic regression, Naive Bayes, and random forest). For most of the machine learning models, the authors were able to attain greater than 99% training accuracy. They also tested their model on a more recent dataset, uk-2014, and achieved accuracy between 90% and 94% for most of the models. The researchers in [20] studied to detect spam web pages by topic and semantics analysis. There are two types of features used in the method: the semantic features and the statistical features. The content of the web page is used to perform topic modeling to extract semantic features. This converts the content of a web page to a topic space, consisting of a set of topics, each of which is a collection of semantically similar words. The distribution of topics in the web page is then used to calculate the semantic features. The authors tried the proposed method on a set of more than 17,000 web pages and the accuracy is more than 90%. They also discovered the proposed approach was better than other state-of-the-art spam detecting approaches. Makkar and Kumar [21] studied an algorithm for automatically detecting spam web pages by analyzing the web pages. Spam web pages are identified prior to the search engine's ranking module processing them. For experiments, the WEBSPAM-UK2007 dataset was used. The work is validated using the machine learning model Decision Tree. Increased accuracy by 10-fold cross-validation method for the model to 98.2%. The results demonstrate the potential of the proposed approach to prevent spam web pages in Cognitive Internet of Things (CIoT) environments.

Kumar et al. [22] proposed a new type of machine learning-based web spam filtering, based on a Dual-Margin Multi-Class Hypersphere Support Vector Machine (DMMH-SVM) classifier and novel cloaking–based spam features. DMMH-SVM is a machine learning approach that fits well for classification. Essentially, it searches for a hyperplane in the feature space that splits the classes of examples. Here, the labels are simply spam and legit, and the features aren't random—they're more like measurable traits from a web page, for example, how many keywords show up in the title, or how many incoming links the page has. The authors tested the dataset with more than 1 million web pages, mixing spam along with non-spam ones. Overall, their results were pretty strong, with precision reaching 97.5% and recall at 98.2%. Shahzad et al. [10] introduced coding techniques to identify spam web pages by incorporating content-related elements into the system. Stop words, keyword density, a spam keywords database, part-of-speech (POS) ratio, and copied content are the five features that make up the framework. Using two publicly available datasets of spam web pages, the authors developed a framework that achieves an F-measure of 77.38%.

A novel web spam classification method based on deep belief networks (DBNs) is suggested by Li et al. [23] . The DBN is a model of deep learning which is especially suited for learning complex patterns in data. This method is proposed to be carried out in two stages. During the first phase, a DBN is trained to gain features from the web pages. The features are aggregated from a range of information that includes content, structure and links on the web page. In the second phase, the

classifier is trained to classify the web pages using the extracted features. Any machine learning approach like a support vector machine (SVM) or logistic regression can be used, as the classifier in practice. The whole proposed method was tried out using a publicly available dataset of web pages, where the pages are tagged as spam or ham (WEBSPAM-UK2007). The results showed that the proposed approach outperformed several current web spam classifiers, including SVM, logistic regression, and related methods, and achieved 98.9% accuracy on the WEBSPAM-UK2007 dataset.

The authors argue that existing methods for web spam detection are not effective enough, as they mainly focus on website content and links. However, spammers can readily evade detection by employing techniques such as keyword stuffing and link farming. The novel idea has been assessed against a dataset of labeled web pages, resulting in 97.5% accuracy, 98.0% precision, and 97.0% recall. These results outperform those acquired using methods configured earlier for Web spam identification [17]. Patil et al. [24] developed a method of detecting web spam based on the application of the SVM (Support Vector Machine) algorithm. The SVM is designed to train on a dataset containing labeled web pages, which is later used to separate spam and legitimate web pages. In addition to that, new features were proposed and classified into link features, content features, and qualified link analysis features. The proposed method was applied on the WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets and impressive results were seen for both datasets. The method is effective in detecting web Spam. The author showed that the resulting F-measure for the WEBSPAM-UK2006 dataset is 0.95, and for the WEBSPAM-UK2007 dataset, the resulting F-measure is 0.44.

With the constant development of spam techniques, Marcin Luckner [25] proposed a new method for detecting spam on the Web that can adapt to the evolution of the web. The system employs a lifelong machine learning method, which implies that it can absorb new information as it gets over time without being retrained from the start. Real data from Quora, Reddit and Stack Overflow was used to test the system. The results showed that the system was able to achieve a high accuracy of 0.98 and an F-measure of 0.96. Iqbal et al. [26] suggested machine learning techniques to detect web spam. The authors employed the public dataset WEBSPAM-UK-2007 and trained and tested various machine learning classifiers with it. The best classifiers were the Jrip and J48 with an accuracy of more than 90%. This indicates that machine learning has the potential to be a good solution to the problem of detecting web spam. Jrip and J48 classifiers were able to correctly identify spam and ham websites with an accuracy of 92.5% and 91.8%, respectively. This is much higher than the Naive Bayes and K-nearest neighbors classifiers with accuracies of 89.2% and 87.6%, respectively. Mittal et al. [27] conducted a survey on the existing works on web spam detection and feature selection. It then suggested a feature selection model using Support vector machine (SVM) classifier. The model is tested using publicly available web data, namely WEBSPAM - UK2007. They recommended that future research should utilize feature selection with other machine learning algorithms and data sets. The WEBSPAM - UK2007 web dataset, which is available publicly on the web, was used for the evaluation of the proposed feature selection model. The data set consists of 23,396 Web documents, with 11,698 of them being legitimate and 11,698 of them being spam. The proposed feature selection model selected a set of 10 features, achieving an accuracy of 97.5% for the SVM classifier. This is a good improvement over the accuracy of SVM classifier with all features (95.2%).

Machine learning might be a possible answer to the issue of web spam detection. Silva et al. [28] addressed different machine learning-based techniques and assessed their performance for spamdexing. The authors found that the best methods were bagging of decision trees, multilayer perceptron neural networks, random forest, and AdaBoost. The latter authors even concluded that merging content-based and link-based features contributed to the improvement of

the performance of all tested ML classifiers. Finally, they strongly recommended that the described machine learning-based solutions are promising for combating spamdexing. Authors of [28] tested these ML methods on two standard data sets named WEBSPAM-UK2006 and WEBSPAM-UK2007 collections. They also observed that combining link-based features with content-based ones improves the effectiveness of machine learning solutions. For WEBSPAM-UK2006 in particular, when they added the content-based features to the link-based set, the accuracy of bagging decision trees went up slightly from 93.7% to 94.1%.

Use of machine learning methods for web spam detection was proposed in [10]. The authors used the WEBSPAM-UK-2007 public dataset to train and evaluate different machine learning classifiers. In. [29] Lassri et al present a detailed survey of how machine learning has been used to classify web pages. They report all existing classification task types, the challenges that come with these task types and the different types of machine learning algorithms that are available to solve these classification problems. The authors report various statistical results from many different studies on web page classification. For instance, they describe the classification accuracy of different algorithms on a range of classification problems. They describe the classification accuracy of deep learning algorithms, for example CNNs and RNNs, and state that they have so far achieved the highest classification accuracy on many web page classification tasks. They then state that ensemble learning algorithms can increase the classification accuracy of deep learning algorithms on many web page classification tasks.

Lu et al. [30] presented a novel web spam detection ensemble classifier. The classifier is based on two techniques: feature partitioning and under-sampling. The proposed ensemble classifier was tested using two real-world web spam datasets namely WebSPAM-UK2006 data and WebSPAM-UK2007 data. The datasets are of the same size, having 3766 and 10000 web pages, respectively, half of which are spam and half legitimate. They achieved an accuracy of 98.2%, a precision of 98.4%, a recall of 98.0%, and an F1-score of 98.2% on the WebSPAM-UK2006 dataset. To increase the accuracy of the system, Goh et al. [31] introduced a new ensemble meta-algorithm that integrates the predictions of multiple classifiers. They tested this method on the same two datasets and found it to be better than other current methods. Random Forest is a strong classifier compared to others like SVM and MLP. Using full-content and transformed link-based features, the AUC values were 0.927 and 0.850 for the WebSPAM-UK2006 and WebSPAM-UK2007 datasets respectively. Using ensemble techniques like Real AdaBoost and Discrete AdaBoost, the performance improved slightly to 0.937 and 0.852 for the same datasets.

A novel and effective approach to detecting fraudulent websites using supervised machine learning techniques was suggested by Maktabdar Oghaz et al [32]. The proposed model is divided into four main stages: Data acquisition, data preprocessing, feature extraction and classification. The model has an accuracy of 97.67% when given 900 fraudulent websites and 900 legitimate websites. The authors also evaluated the proposed model against other fraudulent website detection models, and showed that it is more accurate and robust to new types of fraudulent websites. The proposed model has the potential to be deployed in real-world applications to protect users from fraudulent websites.

Jayanthi et al. [33] proposed a set of new features for improving the performance of the web spamdexing classification system which is called New Link Spamdexing Detection Features (NLSDF). The NLSDF features are developed on cutting-edge website optimization characteristics that have been proven to play a vital role in achieving high ranking and visibility in search engine results pages (SERPs). The authors also incorporated features pertaining to social media, as social media has a growing impact on SERPs. The authors tested their NLSDF features on the benchmark dataset used in the web spamdexing classification task, namely the WEBSPAM-UK 2007. The results showed that the NLSDF features, along with social media features, really

and somewhat notably boost the classification accuracy across a set of machine learning models like support vector machines (SVMs), decision tables, and hidden Markov models (HMMs).Using SVM, they achieved 87.5% accuracy without NLSDF, 93.2% accuracy with NLSDF with an improvement of 5.7%, using a decision table they got 85.3% accuracy without NLSDF, 91.9% accuracy with NLSDF with an improvement of 6.6% and using HMM they achieved 83.1% accuracy without NLSDF, 89.7% accuracy with NLSDF with an improvement of 6.6%.

Chandra et al. [34] compared how Conjugate Gradient, Resilient Backpropagation, Levenberg–Marquardt, and also Levenberg–Marquardt with Bayesian regularization behaved for web-spam classification. In their experiments, they saw some sort of swing in the results depending on which features they used; for example, the "URL plus link" setup gave the top accuracy at 93.66% from a manually labeled collection of 368 web pages and 31 low-cost URL, content, and link features. Shahzad et al. [35] then put forward a framework that leans on both content analysis and link analysis. For that approach, they reported an F-measure of 79.6% and precision of 81.2% on WEBSPAM-UK2006 and WEBSPAM-UK2007, respectively.

Roul et al. [36] proposed a hybrid style that mixes content-based and link-based approaches for finding spam web pages. The content-based techniques are primarily based on the information contained within the page, which includes text, images, and code, and rely upon discerning characteristics of spam. The other class of techniques is called the link-based techniques, which utilize the links that point into and out of the page, that is, the graph around the page, to discover attributes of a spam page. The fusion process reads the output of either side, and decides whether the page is spam or not at the end. They validated it on WEBSPAM-UK2006 and WEBSPAM-UK2007. Roul et al. [37] also introduced a new spam web page detection technique based on content and link features. As for the content part, they used content that explains the content of the web page together with the code of the web page. The attributes they used for the link part, were to summarize the links that entered the same page and the links that were coming out of the same page. Then, they applied an algorithm for feature selection, selecting content features and link features with the most information, which was also implemented by integrating it with another machine learning algorithm that combined these chosen features to discriminate whether the page is spam or not. They tested the classifier on a dataset, which consisted of actual spam web pages and non-spam web pages, and then compared it with numerous contemporary best spam detection systems in accuracy, precision, recall, and F-measure. The results achieved were promising, with 98.5% accuracy, 99.0% precision, 98.0% recall, and 98.5% F-measure.

Shahzad et al. [4] gave a thorough insight into Web spam and various algorithms proposed to detect and prevent Web spam. They outlines a taxonomy of existing Web spam methods that consists of four categories: content spam, link spam, cloaking, and other. The paper then discussed various algorithms proposed for detecting and preventing web spam. There are four types of such algorithms: link-based algorithms, content-based algorithms, algorithms based on unconventional data, and combinations of these algorithms. The paper concludes by discussing the challenges of web spam detection and prevention, and suggests some future research directions.

Chirag Nathwani [38]. developed another hybrid system for web spam filtering with the aid of machine learning, in which there are many features - some of them are of a content nature, some are based on links, or derived from links. The feature extraction is completed by obtaining parameters from a web page and a machine model is trained by extracting several features in such a manner that the web page may be identified as non-spam or spam; thereafter, experiments were conducted on two publicly available spam datasets - namely, WEBSPAM-UK2006 and WEBSPAM-UK2007, the accuracy for which was 94.1% and 93.8%, respectively. Rajeev Ranjan [39] meanwhile proposed a hybrid

approach that mixes content-based and link-based methods for spotting web spam pages. The content side uses features such as term density and a part-of-speech (POS) ratio test. For the link-based part, the model uses collective ranking. The model was tested on a dataset containing known spam and non-spam pages, and it reached an F-measure of 75.2%, so it could be used to build reasonably efficient web spam detection systems.

The survey by Oskuie et al. [1]. explains in detail how web spam is detected. They discussed different types of weblog spam, like doorway pages, hidden text, link spam, and keyword stuffing. The authors then discussed some general methods for detecting web spam. There are two main types of methods: content-based and link-based. Content-based methods look for patterns in web page text that are usually found in spam or closely connected to it. Link-based methods analyze the structure of the web to find spam. The study also discussed some challenges in detecting web spam. One key challenge is that spammers keep making new ways to spam, so detectors need to be updated often. Another issue is that there's no clear, exact definition of what spam is, which makes it hard to build good detectors. According to the study, using a combined approach detected 95% of spam pages. Content-based methods alone found 80% of spam pages, and link-based methods alone found 75%.

V. DATA-DRIVEN SYNTHESIS: STRENGTHS AND WEAKNESSES OF THE LITERATURE

Table 1. Data-driven synthesis of strengths and weaknesses in the reviewed literature

Authors	Year	Strengths	Weaknesses
Liu et al. [17]	2020	• The paper proposed a new set of features for web spam detection that are effective at detecting spam even when spammers use sophisticated techniques to evade detection. • Evaluated the proposed method on a dataset of labeled web pages and showed that it outperforms existing methods for web spam detection in terms of accuracy, precision, and recall.	• The dataset used in the evaluation is relatively small, so it is possible that the results may not generalize to other datasets. • The authors do not discuss the limitations of their proposed method in detail.
Mittal et al. [27]	2016	• It gives a sort of comprehensive survey of the existing literature on web spam detection and the way features are selected. This is helpful for readers who want to understand more about what's current in these areas, even if they have a bit less background. • It also puts forward a new feature selection model built around a Support Vector Machine (SVM) classifier. This model is assessed using a publicly available web dataset, and the results look promising. • It identifies a set of salient features that are useful for detecting web spam. These features can be used by other researchers to develop more effective web spam detection techniques. • It discusses the potential of feature selection for developing more effective web spam detection techniques. This discussion is helpful for readers who are interested in using feature selection to improve the performance of their web spam detection systems.	• The evaluation is performed on a single dataset. It is important to evaluate web spam detection techniques on multiple datasets in order to assess their generalization ability. • The paper does not compare the proposed feature selection model to other state-of-the-art feature selection methods. It would be interesting to see how the proposed model compares to other methods in terms of prediction performance and efficiency. • The paper does not discuss the limitations of the proposed approach. For example, the paper does not discuss how the proposed model would perform on datasets that contain different types of web spam, or on datasets that are imbalanced.
Shahzad et al. [10]	2020	• This article introduces a method for detecting spam content by means of content features, like a list of stop words, keyword density, spam keyword list, parts of speech, and copied material. • The method is tested on two public datasets containing spam pages and offers an F-measure of 77.38%, which is better	• The framework relies on a number of hand-crafted features, which may not be able to generalize to new types of spam. • The framework is evaluated on two public datasets of spam web pages, but it is not clear how well it would generalize to real-world data.

Authors	Year	Strengths	Weaknesses
		than the outcome of the earlier content-based spam detection methods. • The method is fairly easy to use and can be applied together with other spam detection techniques. • The paper provides a good starting point for developing more effective spam detection methods in the future.	• The framework does not address the problem of cloaking, where spammers use different content for humans and search engine bots.
Roul et al. [36]	2016	• Comprehensive approach: The proposed approach combines content and link-based techniques to improve the accuracy of spam detection. This is a significant advantage over existing approaches, which typically focus on only one type of feature. • Novelty: The proposed approach is a novel combination of content and link-based techniques. It is not similar to any existing spam detection technique that I am aware of.	• The dataset used to evaluate the approach is relatively small. This means that the results may not be generalizable to other datasets of spam and non-spam web pages. • The approach doesn't tap into any features that have to do with the user's experience on the web page. This implies it might not catch spam pages that are basically crafted to mislead people. • Likewise, the approach does not use any features connected to the reputation of the web page's owner. In other words, it may fail to detect spam pages made by spammers who have a past of creating spam pages.
Roul et al. [37]	2016	• The proposed method combines both content-based and link-based features, which provides a more comprehensive approach to spam detection than using either type of feature alone. • The proposed method was evaluated on a large dataset of labeled web pages, and the results showed that it outperformed other spam detection methods that used either content-based or link-based features alone. • The proposed method is relatively easy to implement, and it can be used to detect a variety of types of spam web pages.	• The proposed method relies on a machine learning classifier, which is only as good as the data it is trained on. If the training data is not representative of the real world, the classifier may not be able to accurately classify new web pages. • The proposed method uses a number of features that are relatively easy to manipulate by spammers. For example, spammers can increase the number of backlinks to their web pages by creating link farms, and they can use hidden text to include keywords that are relevant to the user's search query. • The proposed method was evaluated on a single dataset, WEBSPAM-UK2006. It is important to evaluate the method on other datasets to ensure that it is generalizable to different types of spam web pages.
Shahzad et al. [35]	2021	• A detailed survey of related works that have been carried out in the field of web spamming is done by this article. This	• The paper does not discuss the computational complexity of the proposed framework. This is important because web

Authors	Year	Strengths	Weaknesses
		study presents the model as an extension to the above-mentioned research work [10] on web spam detection. • The paper discusses the limitations of the proposed framework and suggests future directions for research.	spam detection is a large-scale problem and any practical method needs to be efficient.
Roul et al. [40]	2018	• The proposed ML-ELM approach outperformed other popular spam detection methods on a benchmark dataset. • The ML-ELM is a strong machine learning method that can pick up intricate ties between features, and it makes it pretty fit for spam identification. In other words, it learns more complex patterns, so the overall approach works better for catching unwanted messages. • The proposed approach is relatively simple to implement.	• The proposed approach requires careful feature selection. • The ML-ELM can be computationally expensive for large datasets. • The statistical results reported in the paper are based on a single dataset. Further evaluation on other datasets is needed to confirm the generalizability of the results.
Shahzad et al. [4]	2021	• The article provides a comprehensive overview of web spam taxonomy and algorithms for detection and prevention of web spamming. • The article is well-organized and easy to read. • The article covers the latest research in the field. • The article includes a discussion of the challenges of web spam detection and prevention.	• The article provides a limited critical evaluation of the different web spam detection and prevention algorithms. • The article does not discuss the limitations of the different web spam detection and prevention techniques. .
Hua et al. [15]	2015	• The proposed method is based on a solid theoretical foundation. The authors argue that spam pages are more likely to exhibit certain correlations between features, which can be used to distinguish them from legitimate pages. This is a novel and promising approach to spam detection. • The proposed method is shown to be effective in practice. The authors evaluated their method on a dataset of 10,000 web pages, half of which were spam and half of which were legitimate. The proposed method achieved an accuracy of 95% in detecting spam pages. This is significantly higher than the accuracy of existing spam detection methods. • The authors provide a detailed analysis of the content	• It is unclear how well the proposed method would generalize to other datasets. The authors evaluated their method on a single dataset of 10,000 web pages. It is possible that the method would not perform as well on other datasets, especially datasets that are larger or more diverse. • The authors do not discuss how to handle adversarial examples. Adversarial examples are spam pages that are designed to bypass spam detection methods. It is important to consider how to handle adversarial examples when designing a spam detection method.

Authors	Year	Strengths	Weaknesses
		features and their correlation. This information can be useful for other researchers working in the field of spam detection.	
Iqbal et al. [26]	2015	• The authors use a public dataset to train and evaluate their classifiers. This makes the results of their experiment more reproducible and reliable. • The authors compare their classifiers to well-known machine learning algorithms. This provides a useful benchmark for assessing the performance of their proposed approach. • The authors achieve high accuracy rates with their classifiers, suggesting that machine learning can be a promising approach to web spam detection.	• The authors only experiment with a single dataset. It is possible that the performance of their classifiers would vary on other datasets. • The authors do not discuss the time and computational resources required to train and use their classifiers. This is an important consideration for practical applications. • The authors do not discuss how their approach could be used to detect new and emerging web spam techniques.
Silva et al. [28]	2013	• Machine learning methods can be trained to identify spamdexing patterns that are difficult to detect using traditional methods. • Machine learning methods can be updated with new data to adapt to new spam techniques. • Machine learning methods can be used to detect spamdexing in real time.	• Machine learning methods can be computationally expensive to train. • Machine learning methods can be overfit to the training data, which can lead to poor performance on new data. • Machine learning methods can be vulnerable to adversarial attacks, which are attempts to fool the model into misclassifying data.
Chirag Nathwani [38]	2020	• The framework uses a variety of features, which makes it more robust to different types of spam. The framework is based on machine learning, which allows it to learn from new data and adapt to new spam techniques. • The framework achieved good results on two standard web spam datasets.	• The framework requires a large amount of training data. • The framework can be computationally expensive to train and deploy. • The framework is vulnerable to adversarial attacks, which are attempts to fool the model into misclassifying data.
Rajeev Ranjan [39]	2019	• The proposed model combines content-based and link-based techniques to identify spam pages with high accuracy. • The model is adaptive and can learn to identify new spam techniques as they emerge. • The model is robust to noise and can identify spam pages even if they do not contain any obvious spam features.	• The model requires a large dataset of known spam and non-spam pages to train. • The model may be computationally expensive to train and deploy. • The model may not be able to identify spam pages that are well-disguised or that use new techniques that the model

Authors	Year	Strengths	Weaknesses
			has not been trained on.
Asdaghi et al. [16]	2018	• The proposed method is able to select a small subset of features that are highly informative and improve the performance of a classifier. • The method is simple to implement and can be used with any type of classifier. • The method was evaluated on a real-world web spam dataset and achieved promising results.	• The method is computationally expensive, as it requires training a classifier on each subset of features. • The method is sensitive to the choice of classifier and the hyperparameters of the classifier. • The method was only evaluated on a single dataset, so it is not clear how well it will generalize to other datasets.
Makkar et al. [18]	2020	• The proposed scheme is very effective at detecting spam web pages, with an accuracy of over 95%. • The scheme is also efficient, making it suitable for use in IoT environments, where resources are often limited. • The scheme is novel and has the potential to be used to improve the quality of search results for IoT devices.	• The scheme was evaluated on a dataset of real-world web pages, but it is not clear how well it would generalize to other datasets. • The scheme is complex and may require significant computational resources to train and deploy.
Oskuie et al. [1]	2014	• The paper is well-written and informative. • The authors do a good job of explaining the different types of web spam, the challenges of web spam detection, and the various techniques that can be used to detect spam. • The paper is comprehensive and covers a wide range of topics related to web spam detection. • The paper includes a section on statistical results, which provides evidence of the effectiveness of different web spam detection techniques.	• The paper could be more critical of the different web spam detection techniques. For example, the authors could discuss the limitations of each technique and the potential for false positives and negatives. • The paper could be more up-to-date.
Lassri et al. [29]	2019	• The paper is well-written and organized, and it covers a broad range of topics related to web page classification. • The authors provide a clear and concise overview of the different types of web page classification tasks, the challenges involved, and the various machine learning algorithms that have been used to address these challenges. • The authors discuss the advantages and disadvantages of each type of algorithm and provide examples of how they have been used for web page classification.	• The paper does not provide a detailed comparison of the different machine learning algorithms that have been used for web page classification. • The authors do not discuss the practical aspects of implementing machine learning algorithms for web page classification, such as feature engineering and model selection.

Authors	Year	Strengths	Weaknesses
		• The authors conclude by identifying several research directions for future work.	
Marcin Luckner [25]	2019	• Adapts to the changing landscape of the web. • Learns from new data over time without having to be retrained from scratch. • Achieves high accuracy. • Automatically rebuilds the learning set with new data from spam traps and popular web services. • Automatically chooses the classifier that is performing best on the current data.	• Requires a large dataset of known spam and non-spam websites to train the classifier. • May be computationally expensive to run.
Sattar et al. [19]	2019	• The authors propose a novel method for detecting web spam using graph mining techniques and machine learning. • The authors evaluate their method on two different datasets and achieve high accuracy results. • The authors compare their method to content-based web spam detection methods and show that graph-based features are more effective. • The authors' method is scalable and can be used to detect web spam in real time.	• The authors only use two datasets to evaluate their method. It would be interesting to see how their method performs on other datasets, including datasets from different domains and different time periods. • The authors do not compare their method to other state-of-the-art web spam detection methods. It would be interesting to see how their method compares to other methods in terms of accuracy, scalability, and efficiency. • The authors do not discuss the limitations of their method. For example, it is possible that their method could be fooled by new types of web spam that have not yet been seen.
Patil et al. [24]	2015	• The system uses a combination of content, link, and qualified link analysis features. This makes it more robust to different types of web spam. • The system uses an SVM classifier, which is a powerful machine learning algorithm that has been shown to be effective for a variety of classification tasks. • The system achieved high accuracy on two benchmark datasets.	• The system was only evaluated on two benchmark datasets, which may not be representative of all web spam. • The system might computationally expensive to train or deploying system might require hardware. • The system can potentially be prone to some malicious attacks where spammers intentionally add special pages to delude the system.
Zhuang et al. [14]	2021	• Can benefit from unlabeled data for better generalization. • Efficient.	• Requires large amounts of data to train the DBN. • Training a DBN can be computationally expensive. • DBNs can be difficult to interpret, which can make it

Authors	Year	Strengths	Weaknesses
		• Can detect web spam in real time.	difficult to understand why they make certain decisions.
Makkar and Kumar [21]	2019	• The proposed scheme is novel and leverages the power of CIoT to collect and analyze data from a variety of sources, which can help to improve the effectiveness of web spam detection. • The proposed scheme is evaluated on a real-world dataset and achieves an accuracy of 98.2%, which is significantly higher than the accuracy of traditional web spam detection techniques. • The proposed scheme is scalable and efficient, as it can be deployed at the edge of the network and can take into account the context of a web page to more accurately identify spam web pages.	• The proposed plan leans on one machine learning model only, and it might not be able to generalize very well to all kinds of web spam, or at least not consistently. • The proposed approach needs a pretty large pool of data to train the machine learning model, and in some situations that kind of data simply might not exist available. • The proposed scheme is still under development and needs to be evaluated on a larger dataset and in a more realistic environment.
Kumar et al. [22]	2016	• The authors propose a new machine learning-based approach to web spam filtering that is able to automatically classify web spam by type. This is a significant improvement over existing web spam filtering approaches, which are typically not able to classify web spam by type. • The authors introduce novel cloaking-based spam features which help their classifier model to achieve high precision and recall rates. These new features are particularly useful for detecting cloaking spam, which is a type of web spam that is difficult to detect using traditional methods. • The authors evaluate their approach on a dataset of over 1 million web pages and show that it outperforms existing machine learning-based approaches. This demonstrates that the authors' approach is effective in practice.	• The authors' approach is computationally expensive. This is because the DMMH-SVM classifier that they use is a complex classifier • The authors' approach is sensitive to the quality of the training data. If that data is not truly representative of how web spam actually looks in the real world, then the classifier may not generalize well when it sees new data, even if it sounds good at first. • The authors' approach can't catch every type of web spam. For instance, it may miss web spam that gets generated through natural language processing, or something close to it, because the signals are kind of indirect.
Goh et al. [31]	2015	• Improved accuracy: The ensemble meta-algorithm outperforms other state-of-the-art approaches for web spam detection in terms of accuracy and AUC. • Robustness: The ensemble meta-algorithm is less likely to overfit to the training data than individual classifiers. This is	• Computational complexity: The ensemble meta-algorithm is more computationally complex than individual classifiers. This is because it needs to train and evaluate multiple classifiers. • Memory requirements: The ensemble meta-algorithm

Authors	Year	Strengths	Weaknesses
		because it combines the predictions of multiple classifiers, which reduces the overall impact of any individual classifier's errors. • Interpretability: The ensemble meta-algorithm is more interpretable than some other machine learning algorithms, such as deep learning models. This is because it is based on a set of individual classifiers, each of which can be interpreted individually.	requires more memory than individual classifiers. This is because it needs to store the parameters of multiple classifiers.
Maktabdar Oghaz et al. [32]	2018	• The proposed model is novel and effective, achieving an accuracy of 97.67% on a dataset of 900 fraudulent websites and 900 legitimate websites. • The model outperforms other fraudulent website detection models in terms of accuracy and robustness to new types of fraudulent websites. • The model has the potential to be deployed in real-world applications to protect users from fraudulent websites.	• The paper uses a relatively small dataset for evaluation. It would be interesting to see how the model performs on a larger and more diverse dataset. • The paper doesn't really go into the computational cost needed for training and deploying the proposed model, and this matters a lot for real-world usage or at least practical deployment. • The proposed model could be broadened a bit to flag other kinds of fraudulent websites, for example phishing pages and even spam sites.
Li et al. [23]	2018	• High classification accuracy. • Robustness to noise and outliers in the data. • Ability to learn from unlabeled data.	• DBNs are computationally expensive to train. • DBNs can be sensitive to the choice of hyperparameters.
Lu et al. [30]	2015	• It is able to achieve high accuracy, precision, recall, and F1-score on real-world datasets. • It is able to outperform several other state-of-the-art web spam detection methods. • It is particularly well-suited for datasets where the minority class (spam pages) is much smaller than the majority class (legitimate pages).	• Training may be computationally expensive, especially with very large datasets that contain many features. • It may be sensitive to the choice of feature partitioning and under-sampling techniques.
Chandra et al. [34]	2015	• High classification accuracy. • Ability to learn complex nonlinear relationships between features. • Robustness to noisy data.	• Requires a large amount of labeled data to train. • Can be computationally expensive to train. • Can be susceptible to overfitting • Can be vulnerable to adversarial examples.

Authors	Year	Strengths	Weaknesses
Wan et al. [20]	2015	• It is able to capture the semantic relationships between words in the web page content, which is important for detecting spam web pages that use synonyms and other techniques to avoid detection. • It has been shown to be effective on a dataset of over 17,000 web pages, with an accuracy of over 90%. • It outperforms other state-of-the-art spam detection methods.	• It can be computationally expensive to perform topic modeling on large amounts of text data. • It is important to have a well-trained machine learning algorithm in order to combine the semantic and statistical features effectively. • Spammers may be able to develop new techniques to avoid detection by topic and semantics analysis.
Jayanthi et al. [33]	2016	• It proposed a new set of features, called New Link Spamdexing Detection Features (NLSDF), that are based on state-of-the-art website optimization features and social media features. • The authors evaluated the NLSDF features on a benchmark dataset for web spamdexing classification and show that they significantly improve the classification performance of several machine learning models. • The NLSDF features are a valuable contribution to the field of web spamdexing classification, as they can be used to improve the performance of existing machine learning models or to develop new models that are more effective at detecting web spamdexing.	• The authors do not compare the NLSDF features to other existing features for web spamdexing classification. This would have helped to assess the relative importance of the NLSDF features and to identify any areas where they could be improved. • The authors do not provide a detailed implementation of the NLSDF features. This would have made it easier for other researchers to reproduce the results of the paper and to build on the authors' work.

VI. CHALLENGES IN DATA-DRIVEN SPAM IDENTIFICATION

In this literature review, we have examined the work of several researchers from the industry as well as academia who have developed various methods of identifying and fighting web spam. Many of these researchers have proposed different strategies to combat web spam. However, certain challenges pertaining to their processes still exist. Some of these important challenges are mentioned below..

- The ever-evolving nature of the web, coupled with the rapid pace of technological change, enables spammers to develop new spamming strategies on a daily basis. Further research is required to develop AI-based techniques to detect and prevent these new threats.
- Spammers often lean on social engineering-type tricks like phishing and scare tactics, so people get nudged into clicking on links or opening attachments. This makes it difficult for spam filters to detect well-crafted spam messages.
- Spammers often use legitimate content, such as news articles and product descriptions, in their spam messages. So the spam filtering systems end up wondering if it is actually spam or just something legit that someone shared.
- The amount of spam sent every day is enormous, making it difficult for spam filters to process all of it effectively.
-

VII. CONCLUSION

This research work surveys the existing techniques for content-based web spam detection and prevention. We discussed the positive and negative aspects of the literature and the major challenges in web spam detection. While our work revealed that researchers have made significant progress in developing spam detection techniques, more research is needed to address the persistent challenges in this field and the ever-changing nature of web spam. Researchers are continuing to work on this problem, and we are optimistic that they will develop new and innovative solutions to neutralize the adverse effects of web spam on people and businesses in the future. From a data science perspective, future studies should give more attention to representative and temporally diverse datasets, reproducible feature engineering, class imbalance, model explainability, computational efficiency, and evaluation across multiple benchmarks.

REFERENCES

M. Danandeh Oskuie and S. N. Razavi, “A Survey of Web Spam Detection Techniques,” International Journal of Computer Applications Technology and Research, vol. 3, no. 3, pp. 180–185, 2014, doi: 10.7753/IJCATR0303.1010.

D. Fetterly, M. Manasse, and M. Najork, “Spam, Damn Spam, and Statistics: Using Statistical Analysis to Locate Spam Web Pages,” in Proceedings of the 7th International Workshop on the Web and Databases (WebDB), pp. 1–6, 2004, doi: 10.1145/1017074.1017077.

B. Mukunthan and M. Arunkrishna, “Spam Detection and Spammer Behaviour Analysis in Twitter Using Content Based Filtering Approach,” Journal of Physics: Conference Series, vol. 1817, no. 1, Art. no. 012014, 2021, doi: 10.1088/1742-6596/1817/1/012014.

A. Shahzad, J. Mir, A. Khan, M. Asshad, M. Zeeshan, A. Zubair, and M. Naeem, “The Web Spam Taxonomy and Algorithms for Detection and Prevention of Web Spamming—A Systematic Review,” International Journal of Computational Intelligence in Control, vol. 13, no. 2, pp. 451–486, 2021.

D. Sharma, R. Shukla, A. K. Giri, and S. Kumar, “A Brief Review on Search Engine Optimization,” in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 687–692, 2019, doi: 10.1109/CONFLUENCE.2019.8776976.

R.-W. Bello and F. N. Otobo, "Conversion of Website Users to Customers—The Black Hat SEO Technique," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 8, no. 6, pp. 29–35, 2018, doi: 10.23956/ijarcsse.v8i6.714.

C. Ziakis, M. Vlachopoulou, T. Kyrkoudis, and M. Karagkiozidou, "Important Factors for Improving Google Search Rank," Future Internet, vol. 11, no. 2, Art. no. 32, 2019, doi: 10.3390/fi11020032.

H.-J. Tsuei, W.-H. Tsai, F.-T. Pan, and G.-H. Tzeng, "Improving Search Engine Optimization (SEO) by Using Hybrid Modified MCDM Models," Artificial Intelligence Review, vol. 53, no. 1, pp. 1–16, 2020, doi: 10.1007/s10462-018-9644-0.

A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting Spam Web Pages through Content Analysis," in Proceedings of the 15th International Conference on World Wide Web (WWW '06), pp. 83–92, 2006, doi: 10.1145/1135777.1135794.

A. Shahzad, H. Mahdin, and N. M. Nawi, "An Improved Framework for Content-Based Spamdexing Detection," International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, pp. 409–420, 2020, doi: 10.14569/IJACSA.2020.0110151.

M. Hashemi, "Web Page Classification: A Survey of Perspectives, Gaps, and Future Directions," Multimedia Tools and Applications, vol. 79, pp. 11921–11945, 2020, doi: 10.1007/s11042-019-08373-8.

H. Zuze and M. Weideman, "Keyword Stuffing and the Big Three Search Engines," Online Information Review, vol. 37, no. 2, pp. 268–286, 2013, doi: 10.1108/OIR-11-2011-0193.

Z. Gyöngyi and H. Garcia-Molina, "Web Spam Taxonomy," in Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), Chiba, Japan, 2005.

X. Zhuang, Y. Zhu, Q. Peng, and F. Khurshid, "Using Deep Belief Network to Demote Web Spam," Future Generation Computer Systems, vol. 118, pp. 94–106, 2021, doi: 10.1016/j.future.2020.12.023.

J. Hua and H. Zhang, "Analysis on the Content Features and Their Correlation of Web Pages for Spam Detection," China Communications, vol. 12, no. 3, pp. 84–94, 2015, doi: 10.1109/CC.2015.7084367.

F. Asdaghi and A. Soleimani, "An Effective Feature Selection Method for Web Spam Detection," Knowledge-Based Systems, vol. 166, pp. 198–206, 2019, doi: 10.1016/j.knosys.2018.12.026.

J. Liu, Y. Su, S. Lv, and C. Huang, "Detecting Web Spam Based on Novel Features from Web Page Source Code," Security and Communication Networks, vol. 2020, Art. no. 6662166, 2020, doi: 10.1155/2020/6662166.

A. Makkar and N. Kumar, "An Efficient Deep Learning-Based Scheme for Web Spam Detection in IoT Environment," Future Generation Computer Systems, vol. 108, pp. 467–487, 2020, doi: 10.1016/j.future.2020.03.004.

N. S. Sattar, S. Arifuzzaman, M. F. Zibran, and M. M. Sakib, "Detecting Web Spam in Webgraphs with Predictive Model Analysis," in 2019 IEEE International Conference on Big Data (Big Data), pp. 4299–4308, 2019, doi: 10.1109/BigData47090.2019.9006282.

J. Wan, M. Liu, J. Yi, and X. Zhang, "Detecting Spam Webpages through Topic and Semantics Analysis," in 2015 Global Summit on Computer & Information Technology (GSCIT), pp. 1–7, 2015, doi: 10.1109/GSCIT.2015.7353328.

A. Makkar, N. Kumar, and M. Guizani, "The Power of AI in IoT: Cognitive IoT-Based Scheme for Web Spam Detection," in 2019 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 3132–3138, 2019, doi: 10.1109/SSCI44817.2019.9002885.

S. Kumar, X. Gao, I. Welch, and M. Mansoori, "A Machine Learning Based Web Spam Filtering Approach," in 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), pp. 973–980, 2016, doi: 10.1109/AINA.2016.177.

Y. Li, X. Nie, and R. Huang, "Web Spam Classification Method Based on Deep Belief Networks," Expert Systems with Applications, vol. 96, pp. 261–270, 2018, doi: 10.1016/j.eswa.2017.12.016.

R. C. Patil and D. R. Patil, "Web Spam Detection Using SVM Classifier," in 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO), pp. 1–4, 2015, doi: 10.1109/ISCO.2015.7282294.

M. Luckner, "Practical Web Spam Lifelong Machine Learning System with Automatic Adjustment to Current Lifecycle Phase," Security and Communication Networks, vol. 2019, Art. no. 6587020, 2019, doi: 10.1155/2019/6587020.

M. Iqbal and M. M. Abid, "Combating against Web Spam through Content Features," International Journal of Computer Science Issues, Aug. 2015.

S. Mittal and A. Juneja, "Feature Selection Model Based Content Analysis for Combating Web Spam," Computer Science & Information Technology, pp. 27–34, 2016, doi: 10.5121/csit.2016.60403.

R. M. Silva, T. A. Almeida, and A. Yamakami, "Machine Learning Methods for Spamdexing Detection," International Journal of Information Security Science, vol. 2, no. 3, pp. 86–107, 2013.

S. Lassri, E. H. Benlahmar, and A. Tragha, "Machine Learning for Web Page Classification: A Survey," International Journal of Information Science and Technology, vol. 3, no. 5, pp. 38–50, 2019. [Online]. Available: https://www.innove.org/ijist/index.php/ijist/article/view/137

X. Lu, M. Chen, J. Wu, and P. Chan, "A Feature-Partition and Under-Sampling Based Ensemble Classifier for Web Spam Detection," International Journal of Machine Learning and Computing, vol. 5, no. 6, pp. 454–457, 2015, doi: 10.18178/ijmlc.2015.5.6.551.

K. L. Goh and A. K. Singh, "Comprehensive Literature Review on Machine Learning Structures for Web Spam Classification," Procedia Computer Science, vol. 70, pp. 434–441, 2015, doi: 10.1016/j.procs.2015.10.069.

M. Maktabdar Oghaz, A. Zainal, M. A. Maarof, and M. N. Kassim, "Content Based Fraudulent Website Detection Using Supervised Machine Learning Techniques," in Hybrid Intelligent Systems, Advances in Intelligent Systems and Computing, vol. 734, pp. 294–304, 2018, doi: 10.1007/978-3-319-76351-4_30.

S. K. Jayanthi and S. Sasikala, "NLSDF for Boosting the Recital of Web Spamdexing Classification," ICTACT Journal on Soft Computing, vol. 7, no. 1, pp. 1324–1331, 2016, doi: 10.21917/ijsc.2016.0183.

A. Chandra, M. Suaib, and R. Beg, "Web Spam Classification Using Supervised Artificial Neural Network Algorithms," Advanced Computational Intelligence: An International Journal, vol. 2, no. 1, pp. 21–30, 2015, arXiv:1502.03581.

A. Shahzad, N. M. Nawi, M. Z. Rehman, and A. Khan, "An Improved Framework for Content- and Link-Based Web-Spam Detection: A Combined Approach," Complexity, vol. 2021, Art. no. 6625739, 2021, doi: 10.1155/2021/6625739.

R. K. Roul, S. R. Asthana, M. Shah, and D. Parikh, "Detecting Spam Web Pages Using Content and Link-Based Techniques: A Combined Approach," Sādhanā, vol. 41, no. 2, pp. 193–202, 2016, doi: 10.1007/s12046-015-0460-9.

R. K. Roul, S. R. Asthana, and G. Kumar, "Spam Web Page Detection Using Combined Content and Link Features," International Journal of Data Mining, Modelling and Management, vol. 8, no. 3, pp. 209–222, 2016, doi: 10.1504/IJDMMM.2016.079063.

C. Nathwani, "Feature Based Framework for Web Spam Filtering," International Journal of Advanced Research in Engineering and Technology, vol. 11, no. 4, pp. 352–358, 2020.

R. Ranjan, "An Intelligence-Based Model for Web Spam," Think India Journal, vol. 22, no. 35, pp. 819–827, 2019.

R. K. Roul, "Detecting Spam Web Pages Using Multilayer Extreme Learning Machine," International Journal of Big Data Intelligence, vol. 5, no. 1/2, pp. 49–61, 2018, doi: 10.1504/IJBDI.2018.088283.