Hindawi
Complexity
Volume 2021, Article ID 6625739, 18 pages
https://doi.org/10.1155/2021/6625739

Review Article

An Improved Framework for Content- and Link-Based Web-Spam Detection: A Combined Approach

Asim Shahzad,[1] **Nazri Mohd Nawi**,[2] **Muhammad Zubair Rehman**,[3] **and Abdullah Khan**[4]

[1]*Faculty of Computer Science, Abbottabad University of Science and Technology, KPK, Abbottabad, Pakistan*
[2]*Soft Computing & Data Mining Centre (SMC), Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor 86400, Malaysia*
[3]*Faculty of Computing and IT, Sohar University, Sohar 311, Oman*
[4]*Institute of Computer Sciences and Information Technology, Faculty of Management and Computer Sciences, University of Agriculture, Peshawar, Pakistan*

Correspondence should be addressed to Abdullah Khan; abdullah_khan@aup.edu.pk

Received 29 July 2021; Revised 25 September 2021; Accepted 15 October 2021; Published 15 November 2021

Academic Editor: Bo Xiao

In this modern era, people utilise the web to share information and to deliver services and products. The information seekers use different search engines (SEs) such as Google, Bing, and Yahoo as tools to search for products, services, and information. However, web spamming is one of the most significant issues encountered by SEs because it dramatically affects the quality of SE results. Web spamming's economic impact is enormous because web spammers index massive free advertising data on SEs to increase the volume of web traffic on a targeted website. Spammers trick an SE into ranking irrelevant web pages higher than relevant web pages in the search engine results pages (SERPs) using different web-spamming techniques. Consequently, these high-ranked unrelated web pages contain insufficient or inappropriate information for the user. To detect the spam web pages, several researchers from industry and academia are working. No efficient technique that is capable of catching all spam web pages on the World Wide Web (WWW) has been presented yet. This research is an attempt to propose an improved framework for content- and link-based web-spam identification. The framework uses stopwords, keywords' frequency, part of speech (POS) ratio, spam keywords database, and copied-content algorithms for content-based web-spam detection. For link-based web-spam detection, we initially exposed the relationship network behind the link-based web spamming and then used the paid-link database, neighbour pages, spam signals, and link-farm algorithms. Finally, we combined all the content- and link-based spam identification algorithms to identify both types of spam. To conduct experiments and to obtain threshold values, WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets were used. A promising F-measure of 79.6% with 81.2% precision shows the applicability and effectiveness of the proposed approach.

1. Introduction

Spamdexing or web spamming is described as "an intentional act intended to trigger illegally favourable importance or relevance for a page, considering the web page's true significance" [1]. Several researchers studied spamdexing issues and found that twenty per cent of web hosts are spam [2]. Web spamming is a well-known challenge for search engines [3]. It massively reduces the SE's results quality. Several users get frustrated while searching for important information when they end up with inappropriate information due to web spamming. As the WWW is growing at an unprecedented rate, the size of available textual data has become huge for any end-user. Worldwide web size's recent survey shows that the web consists of 5.41 billion web pages. Thousands of web pages are being added every day to the web corpus, and many of them are either spam or duplicated [3]. Web spammers use several creative spamming methods for dragging the Internet users to their websites for taking different benefits from them. The goal behind creating the spam pages is to cheat the SE so that it presents spam pages that are entirely nonbeneficial and irrelevant to the web users. The web spammer's ultimate

target is to improve the spam page's rank on a search engine's results page. Besides, there is a substantial economic impact of web spamming; a website with a higher page rank can qualify for a large volume of web traffic and free advertisement. During the past couple of decades, researchers from industry and academia were working hard to propose some advanced techniques for web-spam detection, but web-spamming methods are also evolving and spammers are introducing new spamming methods every day [4]. The research on web-spam detection has become an arms race to challenge an opponent who is consistently introducing new and advanced methods [4]. If one can detect and remove all spam web pages, building a robust and efficient Information Retrieval System (IRS) will be possible. More efficient and advanced SEs, providing results consistent with the user's search query, are currently required. The next significant task would be to order the retrieved web pages by their content or semantic similarity between the search query entered by the user and retrieved pages. The last step would be to present the arranged web pages to the user. Web spamming has several adverse effects on both search engines and end-users [5]. Spamdexing is not only wasting storage space and processing resources but also wasting time. As a search engine needs to index and store many pages, extra storage space is required. Moreover, based on the user's search query, SE needs to search the vast corpus of web pages, so more time is required for searching and retrieving the relevant data. This weakens the search engine's effectiveness and reduces the end-user's confidence in the search engine [6]. Enhancement in antispamdexing techniques is required to overcome web-spam attacks. Any method that is used to obtain an undeservedly high rank for a web page is web spamming. Generally, there are three types of spamdexing techniques: cloaking, link-based spamdexing, and content-based spamdexing. In cloaking, spammers offer content to end-users that is entirely different from the content submitted to the search engine spiders [7]. However, most commonly used web-spamming techniques are content- and link-based web spamming which are investigated in this research work. All techniques used by web spammers to change the logical view that a search engine has over the page content are known as content-based web spamming [1]. It is a widespread type of web spamming [8]. Most of the search engines use information retrieval (IR) models, such as BM25 [9], probabilistic models, statistical-language models [10], and vector-space models [11]. These models are applied to a web page's content for ranking the web page, so content-based web spamming is very popular among web spammers. To manipulate the spam web pages' content, web spammers utilise vulnerabilities of these models [12]. For example, they might use famous keywords many times on a spam web page to increase the keywords' frequencies, copy legitimate website's content, produce the content for spam web pages using machine-generated techniques, and add dictionary's words on a spam web page, giving those words the colour of the background so that these words cannot be seen on the spam web page by the user and are only visible to search engine spiders. These are some of the content-based web-spamming techniques used by web spammers to obtain a higher page rank on search engine's results. Based on the structure of a web page, content-based web spam can be divided into five subcategories: title spam, anchor-text spam, meta-tag spam, body spam, and URL spam. Several other spamming techniques which can target different algorithms used by search engines exist [13]. Another widespread type of spamdexing is link-based spamdexing. Davison [14] described the link-based spamming as the links among several pages that are present for a reason other than merit. To get recognition from link-based algorithms, web spammers build link structures using link-based spamdexing techniques. For example, the PageRank (PR) algorithm will allot a higher page rank to a web page if several other top-ranked websites point to this page with backlinks.

This research work focuses on the detection of both the link- and content-based spamdexing techniques. An improved framework for content- and link-based web-spam detection is proposed. In the proposed approach, stopwords, keywords frequency, part of speech, spam keywords database, and copied-content algorithm are used to detect the content-based web spamming. For the detection of link-based spamdexing, we initially exposed the relationship network behind the link spamming and then used the neighbour pages and spam signals for link-based spam identification. For this purpose, WEBSPAM- UK2006 and WEBSPAM-UK2007 datasets are used. The results with a promising F-measure and better precision show the proposed framework's effectiveness and applicability.

2. Literature Review

In response to the web-spamming challenge [12], different researchers proposed various methods for identifying content-based spamdexing. A group of researchers [3] suggested a few critical content features. They introduced HTML-based characteristics and text compressibility to recognise the content spam. Piskorski et al. [15] explored a considerable number of linguistic features. Latent Dirichlet Allocation (LDA) [16] is widely used for text classification tasks. Biro et al. enhanced the LDA and introduced modified versions, linked LDA [17] and multicorpus LDA [18] models, to detect spamdexing more efficiently. To detect content-based spam web pages, Ntoulas et al. [3] applied the decision-tree classifier. They proposed several features, such as the number of words, the average length of words, the visible portion of the content, and the anchor-text amount within a web page. Another group of researchers introduced the semisupervised method and combinatorial feature fusion [17]. They used semisupervised learning to exploit unlabelled samples and used a combinatorial feature fusion technique to create new features and reduce the term frequency-inverse document frequency (TF-IDF) of content-based features. The results determined the effectiveness of their strategy. Ntoulas et al. [3] did the most fundamental work on the detection of content-based spamdexing algorithms [18–20]. In their research article [20], they suggested the use of statistical analysis. Spam web pages can easily be classified using statistical analysis because web spammers usually create spam pages automatically using the phrase stitching and weaving techniques [1]. These spam web pages

are specifically designed for search engine spiders to obtain a higher rank in search engine results. These web pages are not intended for real humans, so one can observe the abnormal behaviour of these web pages. The researchers also reported that there are a massive number of dots and dashes in the URL of spam web pages and its length is exceptional. During their analysis, they observed that out of hundreds of longest hostnames, eleven were pointing to financial credit-related web pages and eighty were referring to adult content. They also found that these web pages themselves have a duplicating nature; spam web pages hosted by the same host almost contain the same content with minimal variance in the word count. Spammers change the content on these web pages very quickly. They gravely observed the content changing feature for a specific host every week and they tracked the changes on these spam web pages. Finally, they concluded that, based only on content changing features, 97.2 per cent of the most active spam hosts can be detected. Their research identified several other features, which can be seen in [20]. In another research study, these researchers worked on content duplication and concluded that large clusters with similar content are spam [18, 19]. To identify such clusters and duplicate content, they used the shingling technique [21] which is based on Rabin fingerprints [22,23]. Another group of researchers [24] worked on machine-learning models and various other features. They explained how machine-learning models and different features could help in web-spam detection. They achieved the best classification results by using easy to calculate content features, LogitBoost, RandomForest, and several learning models. During this work, they identified computationally demanding and global features; for instance, PageRank (PR) could help very little in quality enhancement. Based on their findings, they claimed that the selection of a proper machine-learning model is critical. Urvoy et al. [25] proposed some features to identify the script- or machine-generated spam web pages. The features are based on the structure of the HTML page. For data preprocessing, these researchers used a unique and nontraditional technique. They removed all the content from the web page and only retained its layout. Instead of examining the content of a web page, they just examined the layout to identify the web page duplication. They used fingerprinting technique [22, 23] followed by clustering and identified groups of spam pages which are structurally near-duplicate. By matching the language models [26] for web pages, a group of researchers [27] proposed a technique for spamdexing detection in blogs.

Web spammers use link-spamming techniques for various reasons ranging from malware propagation to monetary activities. Due to the fast growth and volatile nature of the Internet, spammers are introducing more sophisticated techniques to generate more revenue [28–30]. Unfortunately, these practices have several negative impacts on both search engines and user's experience. For instance, users get annoyed when they cannot find what they are looking for and their systems face possibly high-security threats due to malicious content on these spam web pages. Similarly, spam web pages cause issues for search engines by wasting useful resources of SEs. In the past few decades, researchers developed several new antispamdexing techniques to overcome this issue [31–33]. However, web spammers keep a close eye on antispamdexing techniques and continuously improve their spamming techniques to avoid detection. Usually, search engines use the number of inbound links and their ranking to determine a web page's popularity and reputation. If a large number of popular web pages will link to a web page p, it will get a higher rank in search engine results pages [34]. Though it is an excellent technique to define the page ranking of p, spammers exploit this technique to boost their page rank by increasing the inbound links to it. Web spammers often use the so-called Facebook search engine optimisation (SEO) groups, SEO forums, paid-link services, subreddits, and SEO service providers for this reason even though every search engine provides guidelines to website owners for achieving a good rank [35]. Several researchers from academia and industry also worked on different techniques used for link-based spamdexing detection. Based on these techniques' working mechanisms, all web-spam detection algorithms can be subdivided into five categories. The first category deals with recognising suspicious links, nodes, and their subsequent downweighting [36]. The working mechanism of algorithms in this category identifies weak and suspicious links and then penalises them. Several issues in the HITS algorithm were identified by Bharat et al. [37], for instance, the neighbour-graph topic drift and the dominance of mutually reinforcing relationship. To solve these issues, they proposed a method in which they augmented content analysis with link analysis. Another group of researchers [38] studied the same problem and suggested a technique known as the projection-based technique, and they used this technique for calculating authority scores. They changed the eigenvector in the HITS algorithm to get better results. In the second category, spamdexing detection algorithms deal with the topological relationship between a set of web pages with unknown and known labels by applying various propagation rules for computing the labels of other nodes [36]. One of the earliest algorithms from this group is TrustRank (TR). TR uses personalised PR to propagate the trust from a small seed set of excellent web pages [39]. The third category represents the graph regularisation techniques of link-based spamdexing detection algorithms [40]. The algorithms utilise web graph in this category for smoothing the predicted labels. Some results proved the effectiveness of these algorithms. The fourth group of spamdexing detection algorithms is based on the label refinement on a web graph topology [41]. Usually, the researchers use machine-learning methods for label refinement to classify general problems [42], but several researchers used this method for web-spam detection. Finally, the last category of spamdexing detection algorithms is extracting the link-based features for each node and applying various machine-learning methods for spamdexing detection [43]. As they are supported by most of the link-based techniques, HITS and PR are the most critical ones among all the above algorithms.

Abernethy et al. [40] proposed a combined approach of content- and link-based features for spamdexing detection. Their research offered a WITCH algorithm to detect

spamdexing, and they used a graph regularisation classifier (GRC) with a support-vector machine (SVM). To differentiate legitimate web pages from spam web pages, Egele et al. [44] introduced a new technique and used a J48 decision-tree classifier in their approach. They could detect one spam page out of five spam web pages by decreasing the false positive to zero. Prieto et al. [42] proposed a SAAD (spam analyser and detector) spamdexing detection technique based on a heuristic set. For testing and comparing the SAAD with other existing benchmark techniques, they used two public datasets, Yahoo! and e-mail spam (web-spam corpus). Finally, after getting the results, they announced that their proposed approach could generate a secure client environment and protect the users from attacks. For detecting spam web pages using weight properties, Goh et al. [45] proposed a link-based approach. They defined weight properties as the influence of one web node on the other. They used the WEBSPAM-UK2007 dataset for their experiments, and their results outperformed the benchmark algorithms by up to 6.11 per cent improvement at the page level and 30.5 per cent improvement at the host level. Roul et al. [13] proposed a combined approach of content- and link-based techniques for spamdexing detection. They used part of speech (POS) ratio and term density to detect content-based spam and explored the personalised PR to classify web pages as nonspam or spam. For conducting their experiments, they used the WEBSPAM-UK2006 dataset. Finally, they compared their results with the existing techniques. An excellent F-measure of 75.2 per cent shows the effectiveness of their approach. The search engines' performance for spamdexing detection can be improved by combining content- and link-based spamdexing identification techniques. However, few researchers have worked on it, and very little work has been done using an integrated approach. In our framework, for detecting the spam web pages, we combined both content-based and link-based features. Our results demonstrate the significance of the combined approach and provide better results by obtaining excellent F-measure as compared to existing standard techniques.

3. Content-Based Spamdexing Detection

Content-based web-spamming techniques are easy to use and are favoured among web spammers. To detect content-based web spamming, we proposed an improved framework [46]. To obtain the most appropriate threshold values for conducting experiments, we first preprocessed the data and then proposed and improved several different content-based web-spam-detection algorithms. Each proposed or improved algorithm can detect different types of content-based web-spamming techniques, such as the proposed stopword density technique and an improved POS technique that can detect the machine-generated content. The keyword density technique can identify the keyword stuffing on any web page. The keyword density technique identifies spam web pages based on the number of spam keywords on the page, and the copied-content technique can identify the spam web pages created by copying the content from other useful websites. After proposing several different content-based techniques, we combined all techniques to propose a content-based web- spam-detection framework. The details regarding all of the improved and proposed techniques and the proposed framework can be seen in [46].

4. Revealing the Hidden Relationship behind Link-Spam Network

Detecting newly evolved link-based spamdexing is a continuous process for SEs which is a hidden process from end-users. Usually, in SEO communities, link spammers exchange the links of their web pages to create global link farms. The standard web-spam-detection datasets (WEBSPAM- UK2007 and WEBSPAM-UK2006) are quite old, and to propose an improved link-based spamdexing detection framework, it is essential to identify and understand the currently practised spamdexing techniques. To understand new spamdexing techniques and to reveal the secret relationship behind the link-spam network, we performed some experiments to collect the data through data-collection architecture. Several channels are used by individuals to communicate with each other, and the most common of them are Facebook SEO groups, SEO forums, and subreddits. Individuals also contact SEO service providers to improve their visibility in SERPs. Therefore, using traditional antispamdexing methods, it is impossible to detect these link farms. We studied the spam web pages and the behaviour of web spammers who are using Facebook SEO groups, SEO forums, subreddit, and paid SEO services during our experiments for data collection. We also maintained the database of current spam web pages and revealed the secret link-based spamming networks. For revealing the secret network, we proposed an architecture for data collection that allows us to identify the web pages participating in link farms and to determine the web pages using Facebook SEO groups, SEO forums, and subreddits to improve their rank.

Moreover, we were able to reveal the secret relationship between these web pages. We proposed this architecture to study link spam and kept in mind that we will use this data and knowledge to suggest and improve the web-spam detection framework for link-based web spamming. The basic concept behind our system is that link spammers mutually exchange reciprocal links. More specifically, when a web spammer X wants to increase the inbound links to X's website, X will contact other web spammers, and all of them will mutually exchange backlinks. Usually, web spammers use Facebook SEO groups, SEO forums, SEO subreddits, and some other platforms to communicate with each other for link exchange and other services. They exchange the links in three different ways:

(i) When a web page x links to another web page y and asks for payment or any other benefits instead of the backlink in return, the link exchange is said to be one-way. Usually, paid-link services sell the backlinks using this technique.

(ii) When a web page x links to a web page y and in return y links back to x, the link exchange is said to be two-way. Two-way link exchange is popular among web spammers.

(iii) In the three-way link exchange technique, web page x does not link directly to web page y. Instead of a direct link, a third web page z is used to create a loop. For example, the web page x links to the web page z, and the web page z links to x. Detecting this type of link exchange is more problematic as compared to the first two. By having this information in advance, we can precisely identify all web pages trying to improve their visibility in SERPs using fraudulent techniques. Therefore, we have designed crawlers for link exchange identification, honeypot accounts for obtaining information and data from web spammers in private messages, and databases for storing URLs of all spam web pages. We analysed ten leading Facebook SEO groups, ten SEO forums, five subreddits, and two SEO service providers. After one year, we identified 437,811 web pages involved in link spamming. We discovered several categories of web spammers during our studies and determined that web spammers are using almost all available platforms for increasing their web pages' ranks. They offer backlinks services and provide Facebook likes, Twitter followers, Instagram followers, YouTube subscribers, post boosting, and video views for different social media accounts connected to a specific website for which they want to increase the page rank. Several web spammers offered us their services for approving Google AdSense for spam web pages during our experiments. You can get the cracked version of all SEO software on these platforms and authorise Google AdSense to earn income from Google Ads. After analysing all the spamming categories, we observed two significant types of spammers with unique features. Both the varieties behave differently, so other techniques are required to identify their spam pages. Besides, a more in-depth analysis of the data exposed evident variations in the type of link exchange and helped us understand the secret-relationship network. The secret-relationship network between URLs displayed in public posts and threads is different from the URLs sent through personal messages. In short, our main contributions are as follows:

(1) We gathered the corpus of web-spam links from Facebook SEO groups, SEO forums, subreddits, purchased links, and SEO service providers. We also used the honeypot accounts for harvesting the web-spam links from the web spammer's messages.

(2) We analysed all the links in our corpus. Instead of entirely depending on the data gathered from different platforms, we crawled respective pages to confirm the link exchange.

(3) We revealed several strategies and techniques used by inexperienced and advanced spammers.

(4) We exposed different types of link farms, wheels, and pyramids currently used by web spammers.

(5) We performed an in-depth analysis of spam web pages and identified spam signals in spam pages.

(6) We used these spam signals in our framework for the identification of link-based web spamming.

4.1. Data-Collection Architecture. Figure 1 represents the general structure of our technique. The main components of our architecture are (i) honeypot accounts; (ii) SEO groups, SEO forums, and SEO services; (iii) web page crawlers; and (iv) spam-links and paid-links databases.

Initially, we planned to execute SEO crawlers for harvesting URLs from Facebook groups, SEO forums, and other platforms used in our experiments. However, we manually collected the links from all platforms due to (i) some policy restrictions of platforms, (ii) diverse technologies used by different platforms for implementing their systems, and (iii) difficulty faced by SEO crawlers in differentiating between legitimate and spam URLs in different discussion groups. These were the main reasons for collecting the spam URLs manually and carefully instead of executing SEO crawlers. The spam URLs are manually collected from public posts, author's signatures, and private messages.

4.1.1. Honeypot Accounts. Honeypot accounts are used to attract web spammers and getting secret information from them which is not available publicly. Hence, technically, honeypot accounts are fake accounts created by us for collecting unrevealed data from web spammers. Web spammers only present this kind of information in private messages, so honeypot accounts are advantageous in exposing web spammers.

4.1.2. Spam- and Paid-Links Databases. We created two types of databases: paid-links and spam-links databases. We observed that some spammers are not interested in backlinks and only offer one-way link exchange for money or other benefits. Therefore, we added all such URLs to the paid-links database. We also used several search engines and different keywords to identify the services and URLs that are selling the backlinks. The spam URLs, collected from public threads and private messages, interested in two- and three-way link exchange, were added to the spam-links database.

4.1.3. Web Page Crawler. We used the web page crawler to map the link structure of the secret networks behind the collected URLs. After defining a specific depth for crawling, our web page crawler followed all outgoing links. For more in-depth analysis, it stored the crawled URLs in our spam-links database. We also defined the mechanism to check if a URL is already crawled. During our experiment, we observed that most of the spam web pages link other pages

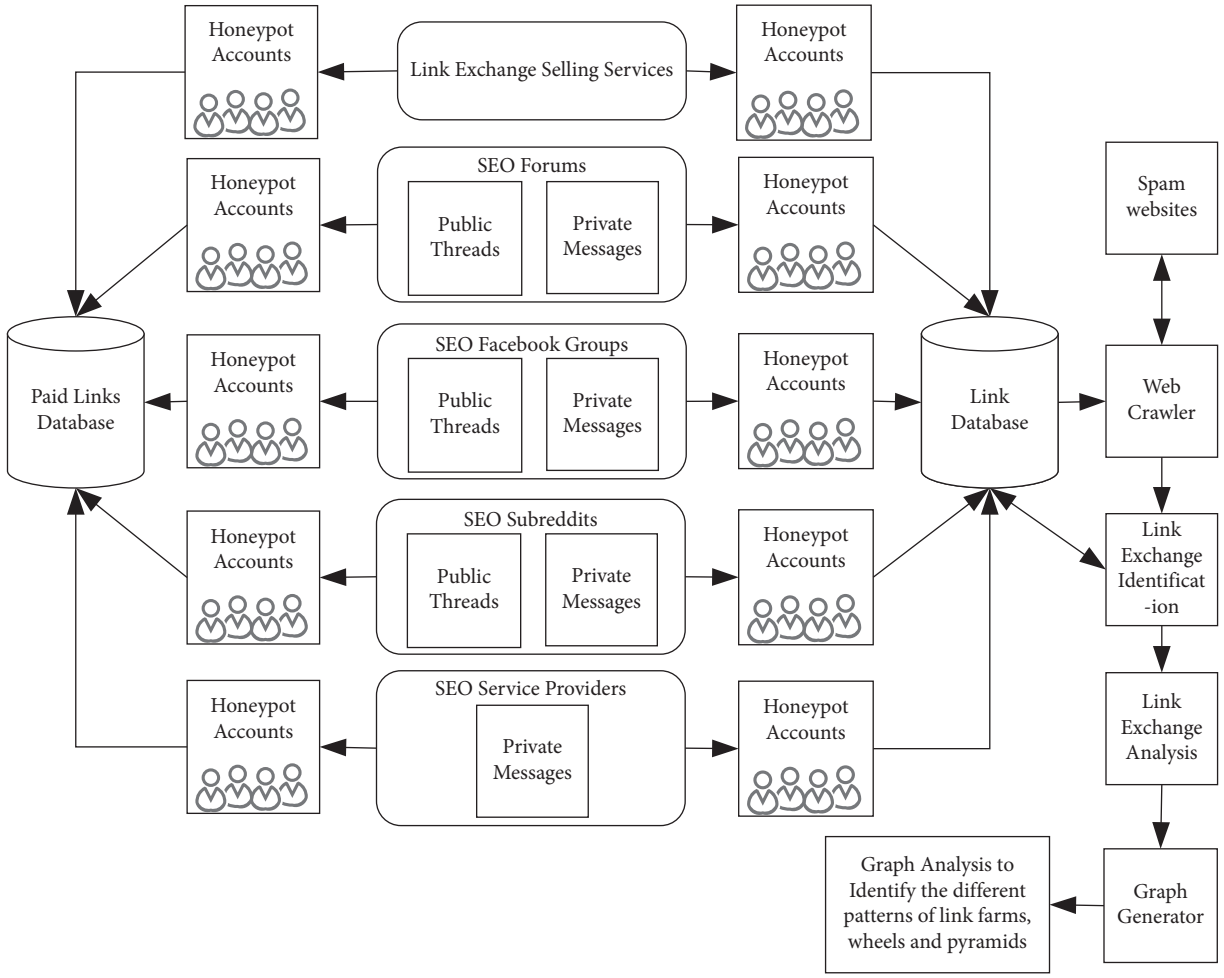

Figure 1: Data-collection architecture for the spam web page.

from their main page. In short, we used a web crawler to validate the link exchange.

4.1.4. Moz Pro. To obtain more accurate results, we used the Moz Pro tool for identifying incoming links of initially collected URLs from public threads, private messages, and SEO services. Moz Pro identified the inbound links, and all the links are stored in spam-links databases. Most web-spam-detection techniques detect spam based on the rule that spam pages usually point to other spam pages. Identifying inbound links will help us in generating a more accurate spam-network graph.

4.1.5. Link Exchange Analysis. After gathering and storing all the necessary information in the databases, we performed the link exchange analysis and correlated the relationships among crawled web pages. More specifically, we identified the relationship between various entities and observed how web spammers interact in link farms. Our approach can recognise all three types of link exchange techniques. We identified different types of link farms, link circles, and link pyramids during our analysis, although three-way link exchange was more difficult to locate.

4.1.6. Web Graph. Representing spam web pages in graphical form can provide a clear view. Therefore, we generated a web graph to visualise the structure of different link farms, link circles, and link pyramids. We recognised significant players on the link exchange and how they create combined link farms, link circles, and link pyramids. All this information helped us in designing the improved framework for the detection of link-based spamdexing.

4.2. Analysis of Facebook Groups, SEO Forums, Subreddits, and SEO Service Providers. For our experiment, we selected the best ten Facebook SEO groups, ten SEO forums, five subreddits, and two SEO service providers and collected the data for one year. We have selected the most active groups, forums, subreddits, and two SEO service providers using black-hat techniques. SEO service providers were carefully chosen after analysing several services and applying the selection criteria. The users use all these places for discussing the SEO boosting methods, so we only target those discussion threads where users were talking about the link exchange and SEO techniques. In total, we analysed 32,674 threads and extracted 137,842 unique URLs posted by 16,827 users and extracted 5,000 unique URLs from the documents

provided by SEO service providers. They provided us with these documents containing backlinks to our website as proof of their work. Our results show that the ratio of replies to each thread is 7.2, which means that for every web page trying to increase its page rank, seven other web pages are available for participation to achieve this target. During this study, we identified two different types of link spammers (beginners and experts). Beginners freely exchange their websites in publicly available posts, while expert spammers only exchange their websites in private messages. Our experiments proved that spammers who exchange their URLs through private messages are more suspicious and are part of bigger link farms and pyramids. Another exciting aspect of understanding the link-based spam network is calculating the page authority, domain authority, and page rank of the website requesting the link exchange. We used SEO PageRank and Moz Pro for calculating the page authority, domain authority, and page rank. After calculating these values, we noticed that most spam web pages have on average the domain authority of seven, page authority of fifteen, and page rank of two, while more than 92% of spam web pages have on average the domain authority of twelve, page authority of seventeen, and page rank of four. This means that the owners of low-ranked web pages behave differently from the owners of high-ranked web pages. Usually, high-ranked web page owners do not exchange links with low-ranked web pages; in some cases, we noticed they are selling the backlinks for money. Every website has a theme of content and the subject of most of the spam pages is earning quick money, finance, travel, adult content, and health. More in-depth analysis revealed that web spammers exchange links with high-ranked or identical-ranked web pages. In public threads, the spammers prefer a two-way exchange while only a small percentage of spammers speak about the one-way or three-way exchanges. On the other hand, in the case of honeypot accounts, most of the spammers are interested in one-way exchange, three-way exchange, and link farms. Initially, through a manual approach, we collected 142,842 unique URLs from honeypot accounts and SEO service providers. For our web crawler, we used these URLs as seeds. In one year, we crawled more than twenty-five million pages, and we were able to identify 437,811 web pages that were engaged in all three types of link exchange. We, by analysing the data, found that the percentage of two-way link exchange is 90.86%, the three-way link exchange percentage is 7.43%, and the one-way link exchange (paid-links) percentage is 1.71%.

5. Data Preprocessing and Experiments for Calculating the Threshold Values for Link-Based Spamdexing

To perform our experiments, we used well-known datasets, namely, WEBSPAM-UK2006 and WEBSPAM-UK2007, and the data was collected using data-collection architecture. To get more accurate results, we manually selected web pages labelled as nonspam/spam by humans. We extracted some link-based features for link-based spamdexing identification. Finally, we obtained the most appropriate threshold values that provided the fewest false positive ratios and high F-measure through different experiments.

5.1. Paid-Links Database. We identified and collected websites that were selling paid links during experiments conducted through data-collection architecture. We also crawled the Internet to search for services that are selling paid links that might increase the ranking and visibility of undeserving websites in search engine result pages. We stored all such services and sites in the paid-links database. We will use this database in our framework for link-based spamdexing detection. Besides, we also stored all spam-web-page links identified through data-collection architecture in another spam-page database. Whenever our framework marks a web page as spam, it will automatically store the spam page's link in the database for reuse in the future. This will help us identify spam web pages and improve the accuracy of our framework.

5.2. Spam-Signal Identification. We manually selected 2,000 spam/nonspam web pages identified during the data-collection experiment and 3,000 web pages labelled as spam/nonspam by humans from the datasets (WEBSPAM-UK2006 and WEBSPAM-UK2007) to analyse web pages for spam signals. We carefully examined all the web pages one by one and identified the following potential spam signals:

(1) Single page website: many spam websites consist of a single web page or very few pages. Though this does not prove that every website with a single or few pages is spam, several spam websites have one or a few pages; hence, there is a correlation.

(2) Thin content: web pages having no or very little content are known as thin content web pages. These pages have no or little value to the user. Usually, these pages consist of copied, machine-generated, or low-quality affiliate content.

(3) No contact information: almost all the spam websites do not have any contact information on their pages. These websites rarely have real phone numbers, e-mail addresses, or any other contact information on their pages.

(4) Presence of spammy keywords: spam sites regularly use particular words directly linked to spam topics like earning money from home, getting free cash, gaming, pharmaceuticals, adult content, and others.

(5) No SSL certificates: most of the spam web pages did not invest in SSL certificates, though HTTPS is an excellent trust signal.

(6) No links to social media accounts: usually, spammers do not associate social media pages with their spam web pages. During our analysis, we found that nearly all spam sites had not associated LinkedIn pages, and Facebook tracking pixel was rarely present on these spam pages.

(7) External outgoing link: the outgoing link points to an external or targeted domain and is different from the link present on the source domain. Spam sites host a massive number of nonrelevant external outgoing links.

(8) Content to links ratio: on legitimate sites, the ratio of the content and external links is balanced, while spam sites have abnormal ratios of content to links. We found that some spam pages consist of external links only and there was no content at all.

(9) The ratio of incoming links: a widespread perception is that only legitimate sites backlink other legitimate sites. In spam sites, we noticed that the ratio of spam incoming links to legitimate incoming links is higher, which is a significant spam signal.

(10) External links in navigation: spam sites host and, in most cases, try to hide many external links in toolbars, footers, and sidebars.

(11) A few internal links: during our page analysis, we found that spam sites have very few internal links. Usually, legitimate sites heavily link to themselves internally to correlate the content. The absence of internal links and navigation is one of the spam signals.

(12) URL length: spam pages character count (URL length) is abnormally long and higher than average size. For example, https://getcheapmedicines.freeshipping.cheappharmacy.com shows the keyword stuffing.

(13) Numerals in domain name: most spam websites contain numbers in their domain names. There is a strong possibility that domain names with numbers in them are generated automatically; therefore, it is a spam signal.

(14) Top-level domains (TLD): some of the top-level domains are correlated with spam domains, for instance, .cc, .pw, .pl, etc. Several spam sites use these TLDs.

(15) Huge proportion of anchor text: in spam sites, the content text proportion is minimal compared to anchor text. A large amount of anchor text in a website is a spam signal.

(16) Site markup proportion: in spam sites, the markup proportion is abnormally small compared to JavaScript and HTML where the proportion of visible text is higher. Typically, legitimate sites invest in a rich user experience with extensive markup CSS and JavaScript.

(17) Broken links: typically, spammers do not maintain and update their websites. We found a lot of broken links in spam pages, which is a spam signal.

(18) Favicon: in today's world, the favicon is considered an essential part of any website, and legitimate websites do use the favicon. Most spam websites do not use the favicon because they change their websites very quickly after getting benefits from it. Missing favicon is a spam signal.

(19) Page not found 404 error: during our analysis, we frequently saw the 404 error on spam sites and the reason behind this is that spammers change their websites very often and they add and remove pages daily, but they do not maintain the site properly. Usually, they do not design a customised 404 error page. The occurrence of too many 404 errors on a website is a sign of a spam site.

(20) Meta description length: standard meta description length is 44–164 characters, while spam sites contain a too short or very long meta description, which is a potential spam signal.

(21) Length of the title: the recommended length of a title is 65–70 characters and spaces are included in it, while spam sites use too short or very long titles, which is a spam signal.

After identifying the spam signals above, we performed a more in-depth analysis to determine spam signals' effects on any website. We randomly selected 5,000 nonspam/spam pages from the datasets WEBSPAM-UK2006 and WEBSPAM-UK2007, and the data was collected using data-collection architecture. To get more accurate results, we manually selected web pages labelled as nonspam/spam by humans. Then, we crawled the URLs and removed all the URLs that did not return a 200 OK. Each spam signal is a potential warning which indicates that a website might be spammy. If a website contains more spam signals, this increases the probability that a website is a spam site, so the total number of spam signals on a site is a powerful predictor of spam. Therefore, we used spam signals to predict a page because several legitimate sites might have one or a few spam signals. We analysed the pages selected for this experiment to show the relationship between the number of spam signals and spam sites. Table 1 shows all the spam signals with the odds ratio for each spam signal. For every spam signal, we can calculate two different percentages: first, the percentage of sites that contain the spam signal and are marked as spam; second, the percentage of sites that contain spam signals and are not marked as spam. The odds ratio increases the possibility that if a site has a spam signal, it is spam. For instance, the first row of Table 1 says that the site with a single page spam signal is 5.2 times more likely to be spam than the site without this spam signal.

The overall spam score is an aggregate of twenty-one different spam signals. Table 2 shows the relationship between the number of spam signals and the percentage of sites containing those spam signals that humans labelled as spam. Danger levels are divided into four different categories: low (<10%), moderate (11–50%), borderline (51–89%), and spam (≥90%).

6. Link-Based Spamdexing Detection

In this section, we present an improved framework for link-based spamdexing detection. Figure 2 shows the flowchart of the proposed framework. This framework uses three different methods, and every technique uses a unique feature to identify link-based spamdexing. The training process is not required as no machine-learning technique is used in the proposed framework. The methods are as follows:

Table 1: Percentage of different spam signals.

Spam signals	Spam vs. nonspam ratio (odds ratio)	Percentage of sites with spam signals and marked as spam (%)	Percentage of sites with spam signals and marked as nonspam (%)
Single page website	5.2	18.87	5.39
Thin content	4.08	26.67	5.09
No contact information	6.78	19.28	2.07
Presence of spammy keywords	11.32	23.59	2.32
No SSL certificates	13.01	27.30	3.83
No links to social media accounts	9.43	36.18	10.98
External outgoing links	4.60	21.47	6.75
Content to links ratio	3.78	36.48	5.89
The ratio of incoming links	9.17	20.06	1.96
External links in navigation	12.52	16.08	5.89
A few internal links	3.06	32.76	6.60
URL length	6.90	21.69	3.63
Numerals in domain name	7.36	35.44	4.69
Top-level domains	11.17	21.29	11.18
Huge proportion of anchor text	2.15	26.78	7.66
Site markup proportion	4.41	10.36	8.87
Broken links	1.19	17.50	7.96
Favicon	2.45	11.37	5.09
Page not found 404 error	2.39	15.27	9.47
Meta description length	1.72	13.78	5.59
Length of the title	2.70	10.70	3.88

Table 2: Spam score based on the number of spam signals in a website.

Number of spam signals	Probability of spam (%)
0	0.68
1	1.10
2	2.50
3	6.98
4	8.10
5	13.04
6	19.09
7	26.89
8	33.79
9	34.68
10	54.05
11	63.20
12	74.77
13	91.75.00
14	100.00
15	100.00
16	100.00
17	100.00
18	100.00
19	100.00
20	100.00

(1) link-based spamdexing detection using paid-links database

(2) Link-based spamdexing detection using spam signals

(3) Link-farm detection using backlinks

All three techniques are discussed in detail in the following sections.

6.1. Link-Based Spamdexing Detection Using Paid-Links Database. Backlinks play an essential role in ranking a website high in search engine result pages, and they also have a high impact on search engine ranking algorithms. That is why backlinks to a website are essential. Web spammers use different techniques to build these links and pay for the backlinks to third-party domains. Paying for the backlinks is strictly banned by search engines. This kind of web spamming can be detected by using the paid-links database. A web crawler is used to identify all backlinks of a web page W*pi*. After identifying all the backlinks, it will compare each backlink of W*pi* with links present in the paid-links database. As paid links are strictly banned and considered spam, a web page with backlinks found to match any link in the paid-links database will be marked as paid-link spam. Otherwise, the page W*pi* will be tested by the spam signals module. Figure 3 shows the algorithm for paid-link spamdexing detection using the paid-link database.

6.2. Link-Based Spamdexing Detection Using Spam Signals. Every single spam signal is a potential warning which might indicate that a website is spammy. Therefore, if the number of spam signals is high on a web page, there are higher

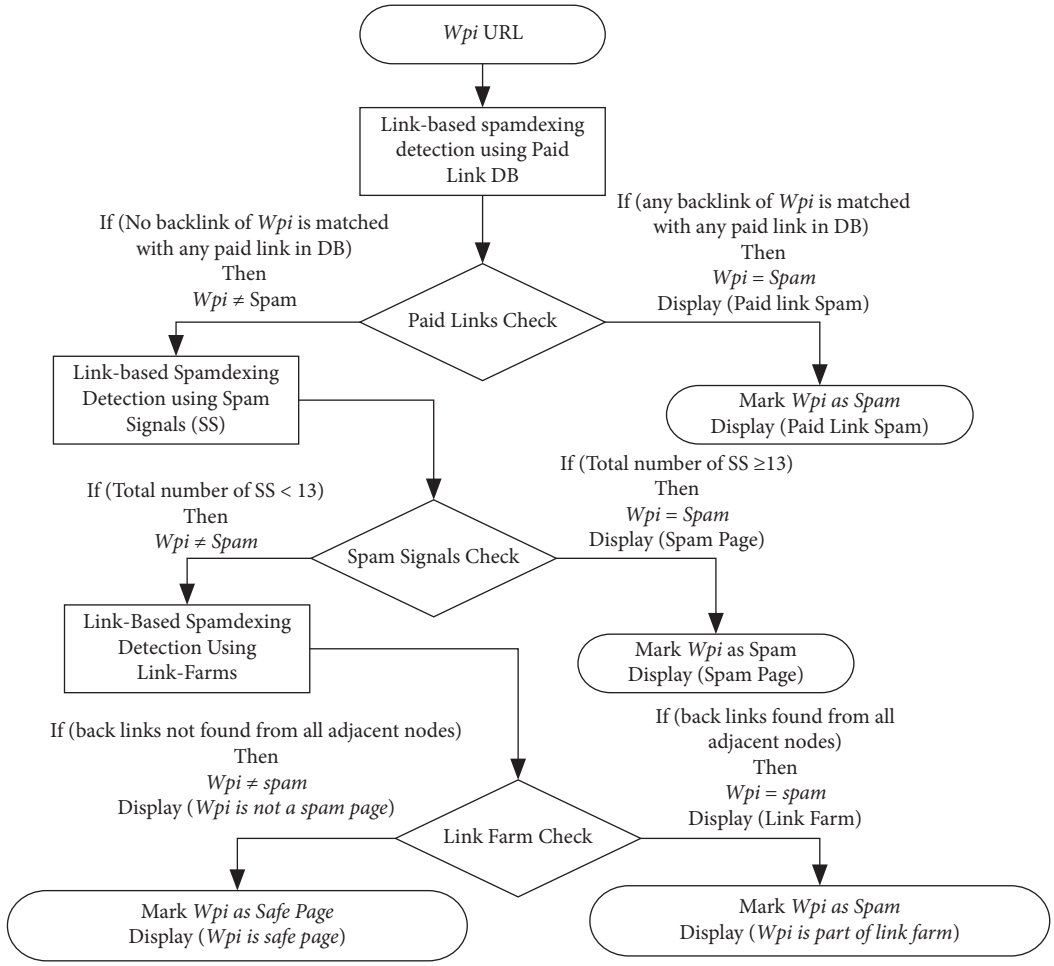

Figure 2: Improved framework for link-based spamdexing detection.

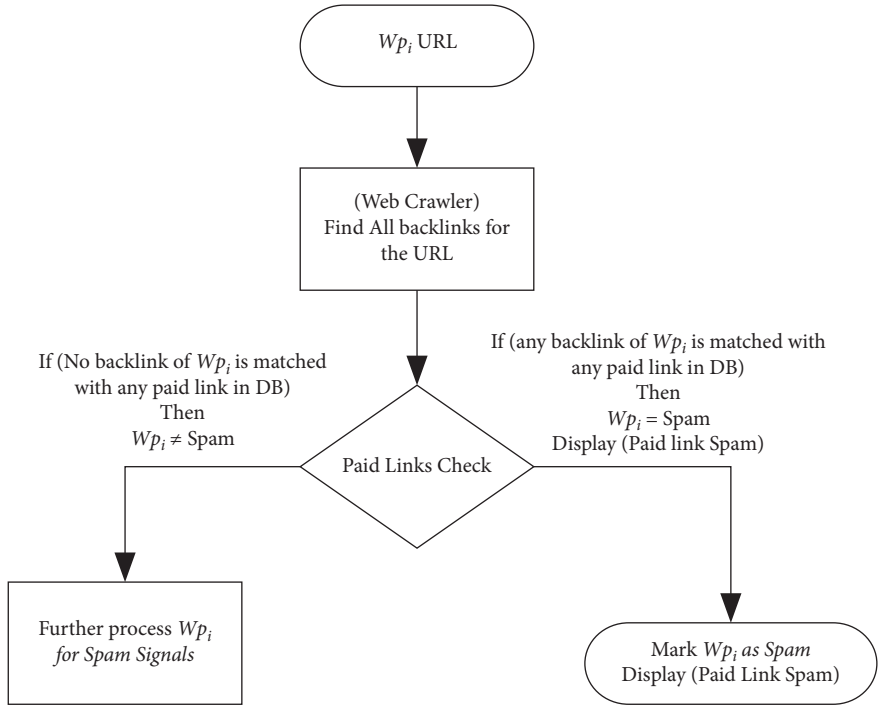

Figure 3: Algorithm for link-based spamdexing detection using the paid-link database.

chances that the page is a spam page. During our analysis and experiments described in Section 5, we identified twenty-one different spam signals and also found the effects of each spam signal on a web page. Our results in Section 5 showed that if the number of spam signals on a web page W*pi* is equal to or greater than thirteen, the web page is spam. Moreover, if the number of spam signals is less than thirteen, there are chances that the page might not be spam, so we need to run further link-based spam tests on all such pages. Since outliers always exist in our experiments, we observed some web pages with very few spam signals but nonetheless marked as spam. There were some web pages with a high number of spam signals on them, but they were marked as nonspam. We have designed a custom web crawler that can identify spam signals on any web page W*pi*. The working mechanism of the spam-signal-detection module for spam-page identification is described as follows:

(1) The spam-signal module will obtain URL of the page W*pi* from the previous module and then submit it to the web crawler.

(2) The web crawler will analyse the page W*pi* and will identify the spam signals.

(3) After the spam-signal identification, it will count spam signals on the web page W*pi*.

(4) Finally, the spam-signal module will check if the total number of spam signals on the page W*pi* is equal to or greater than thirteen. If the number of spam signals is equal to or greater than thirteen, it will mark the page as spam; otherwise, it will forward the URL of the page W*pi* to the next module for further link-based spam-detection test. Figure 4 shows the algorithm for spam signals detection using a web crawler.

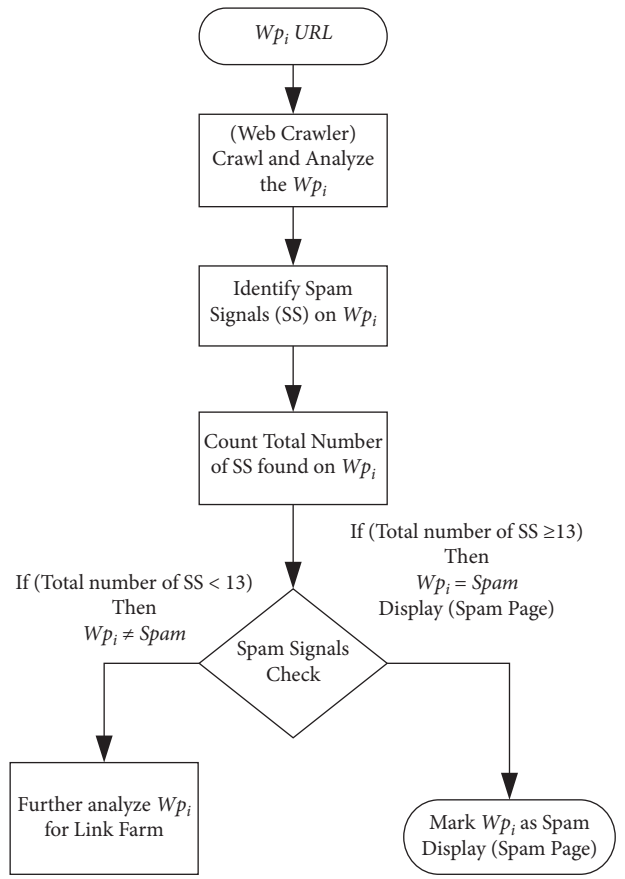

Figure 4: Algorithm for link-based spamdexing detection using spam signals.

6.3. Link-Based Spamdexing Detection Using Link Farm. Any group of websites on the Internet that hyperlink to every other website in the group is known as a link farm. A link farm is a clique in graph-theoretic terms. Although some web spammers create link farms manually, many use automated tools and services to create link farms. Web spammers use private-blog networks and link farms to boost ranks of their websites.

The working mechanism of the link-farm-identification module is as follows:

(1) The link-farm module will obtain the URL of the web page W*pi* from the previous module and submit it to the web crawler

(2) The web crawler will crawl all the adjacent nodes (outgoing links) of the web page W*pi*

(3) The web crawler will crawl adjacent nodes one by one to the defined depth

As W*pi* and all its adjacent nodes are crawled, a web graph is obtained and this web graph can be represented with mathematical notations as shown below. Figure 5 shows the algorithm for link-based web-spam detection using link farm.

Let us say $G = (V, E)$ is a directed graph, V represents the nodes or vertices, and E represents the directed edges. G can be described as a web graph in which V denotes the set of web pages and E represents a set of links between pages. Moreover, it is worth mentioning that if there are several incoming or outgoing links between two pages, our algorithm will consider only one link as proof of a connection between two pages. We used the following notations and definitions in this research.

The out-degree *(od)* of a web page W*pi* is the total number of outgoing links:

$$od(Wpi) = \sum_{Wpi} E_{ij}. \quad (1)$$

The in-degree (id) of a web page W*pi* is the total number of inbound links:

$$id(Wpi) = \sum_{Wpi} E_{ji}. \quad (2)$$

The formula in (1) and (2) can be used to create an adjacency matrix. An element Aij = 1 if web page W*pi* has a link to web page W*pj*; otherwise, Aij = 0. We can call it the link/connectivity matrix.

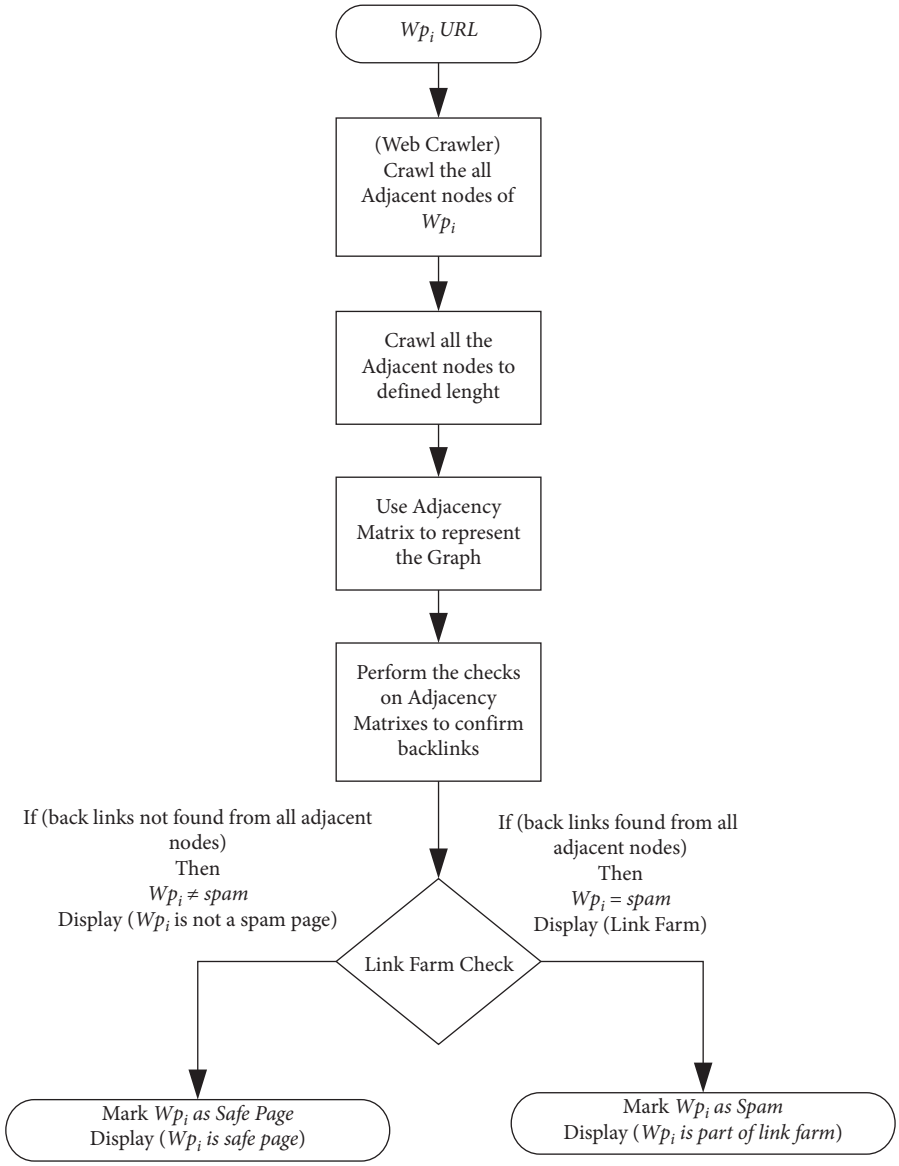

Figure 5: Algorithm for link-based spamdexing detection using link farm.

$$A_{ij} = \begin{cases} 1, & \text{if}\left(Wp_i, Wp_j\right) \in E, \\ 0, & \text{Otherwise.} \end{cases} \tag{3}$$

For a directed graph, the generalised n ч n adjacency matrix A is shown as follows:

$$A = \begin{pmatrix} a_{1,1} & a_{2,1} & \cdots & a_{n,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n} & a_{2,n} & \cdots & a_{n,n} \end{pmatrix}. \tag{4}$$

For checking reciprocal links between two web pages, the algorithm will follow the following procedures. The procedures for link-farm identification are explained with the help of the diagram in Figure 6. Consider that our web crawler identified all the adjacent nodes and found all the reciprocal links between the web page Wpi and all its neighbouring nodes so that a web graph, shown in Figure 6, is obtained.

Initially, it will create the adjacency matrix. If adjacent nodes also have adjacent nodes other than the nodes adjacent to Wpi, they will be ignored and will not be added to the adjacency matrix. For instance, in the web graph in Figure 6, Wpi has *A*, *B*, *C*, *and D* adjacent nodes, so only these nodes will be considered for the adjacency matrix; the nodes adjacent to *A*, namely, *H* and *G*, will be ignored. Similarly, if a web page links back to itself, it will also be ignored and will be represented with *0* in the adjacency matrix. For instance, in the example above, node *B* is linking back to itself and is, therefore, ignored. Accordingly, the following will be the adjacency matrix for the web graph in Figure 6.

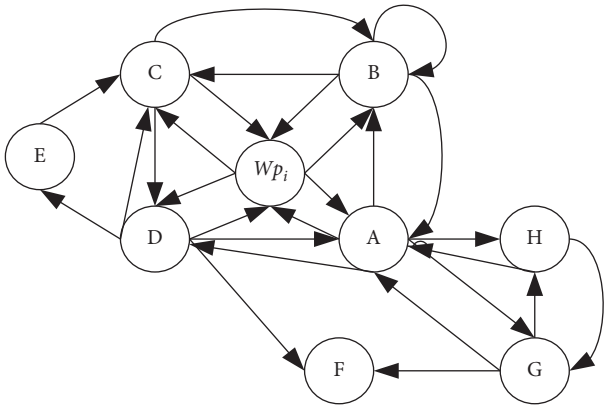

Figure 6: Web graph representing the websites and links.

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}. \quad (5)$$

If $a_{1,2} = a_{2,1} = 1$, this means there exists a reciprocal link between W*p1* and W*p2*; 1 in the adjacency matrix represents the backlink and zeros represent no link. To confirm that every web page in the group links to all other pages of the group, the program will replace all 0s with 1s and all 1s with 0s in the adjacency matrix. The following will be the new adjacency matrix.

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \quad (6)$$

Finally, the resultant adjacency matrix will be checked. If the resulting matrix is an identity matrix, each web page in the group links to every other page in the group, and W*pi* is a member of a link farm and will be marked as spam. If the resultant matrix is not an identity matrix, the link-farm module will mark every adjacent node of W*pi* as W*p2* W*p3*,..., W*pn* and will apply the same link-farm detection technique on every W*pn*. If any adjacent node W*pn* is a part of a link farm, W*pi* is not directly connected to any link farm, but W*pi* is a part of the link pyramid. Wpi will be marked as spam.

7. Combined Approach for Content- and Link-Based Spamdexing Detection

In this section, we discuss a combined approach used for the identification of spam web pages. We combined the content-based framework with a link-based framework so that the final output of the content-based approach will be the input of the link-based approach. During our experiments and analysis, we observed that web spammers are practising different spamdexing strategies. Some spammers only focused on content-based spamdexing, some targeted the link-based spamdexing techniques, and several were involved in both types of spamdexing at the same time. Therefore, the combined approach for spamdexing detection is better in that it can check any web page for both techniques of spamdexing. It is possible that a page is involved in content-based spamdexing but does not involve link-based spamdexing techniques and vice versa. For instance, if a technique is designed for content-based spamdexing detection, the technique can only detect the pages involving content-based spamdexing. It will mark the web page as clean, which involves link-based spamdexing. Unfortunately, several researchers in the field focused on a single detection technique only, and a few worked on a combined approach. The working mechanism of the combined approach is described as follows:

(1) The combined approach will accept a web page as input and check the web page for content-based spamdexing using the five different methods discussed in Section 6. If the web page involves content-based spamdexing, it will mark the page as spam and the process will stop there. However, if the web page passed all the content-based spamdexing identification checks, the URL of the web page will be forwarded to the link-based spamdexing detection section for further analysis.

(2) There are possibilities that a web page does not involve content-based spamdexing but involves link-based spamdexing only. Now, whether the web page is engaged in link-based spamdexing will be checked by scanning it using link-based spamdexing methods. If the page is involved in link spamming, then it will be marked as spam; otherwise, it will be marked as nonspam. Figure 7 shows the complete framework of the proposed combined approach.

8. Results

To conduct experiments, our verification set consists of randomly chosen web pages labelled as spam and nonspam. These web pages are selected from the dataset obtained through experiments in Section 4, using WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets. These datasets are well known and are most suitable for web-spam detection due to the following properties:

(1) The datasets are a mixture of spam and nonspam web pages practising several different web-spamming techniques.

(2) All the researchers in the field can freely access the datasets from the official website and use them as a benchmark measure to detect spam web pages.

(3) The sample web pages in the datasets are random and uniform.

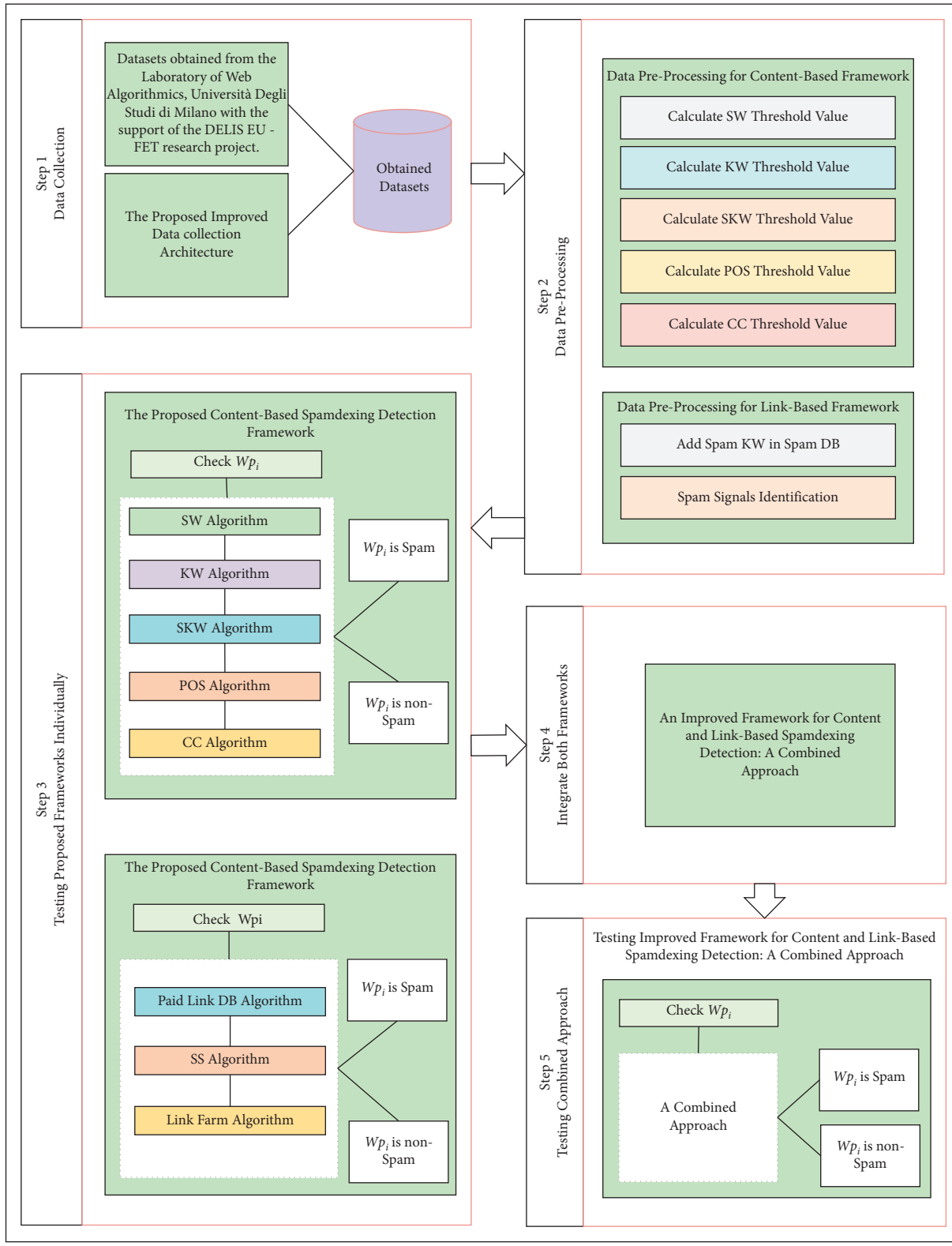

Figure 7: The proposed framework for the combined approach.

(4) The datasets consist of several different kinds of spam pages created using many types of web-spamming techniques.

(5) The web pages are split into training and testing sets with both spam and nonspam labels. Therefore, these datasets can be used to perform experiments and to

detect both content- and link-based web-spamming techniques.

(6) We used these datasets to obtain the optimised threshold values for our proposed combined approach.

There are 11,402 hosts in WEBSPAM-UK2006, but only 7,473 are labelled, while WEBSPAM-UK2007 contains 114,529 hosts in total, and 6,479 are labelled. The following preprocessing steps are performed to obtain a dataset of five thousand web pages:

(1) The web pages labelled as spam or nonspam by real humans are considered

(2) We further filtered out the human label pages and only considered the web pages that are currently working/existing links

(3) Among the currently existing web pages, we only considered web pages having at least one KB of content

(4) Furthermore, the content of these web pages is extracted and stored in a text file format

To implement the proposed combined approach, Python was used, and a machine with 128 GB DDR3, 2x Intel Xeon E5-2670 V2 2.5 GHz 10 Core, and operating system Ubuntu 14.04 was used for the execution of algorithms. As F-measure is a standard approach for combining precision and recall, we used the F-measure to compare our work with other similar related works and to evaluate the proposed framework. Our proposed combined approach for spamdexing detection achieved the results shown in Table 3.

Table 3: Performance evaluation of improved framework.

Technique	Precision (%)	Recall (%)	F-measure (%)
Content-based	78.3	75.6	77
Link-based	73.5	69.7	71.5
Combined	81.2	78	79.6

Table 4: Comparison of the proposed combined approach with other standard techniques.

Combined techniques	Precision	Recall	F-measure
The proposed framework	81.2	78	79.6
Roul et al.	72.9	77.6	75.2
Dia et al.	65	44.3	52.7
Benczúr et al.	67.1	76.7	71.6
Egele et al.	51.2	35.6	41.9
Becchetti et al.	68.8	76.2	72.3

9. Comparison with Existing Approaches

We compared our results with the following existing techniques. The comparison results in Table 4 clearly show that the proposed combined approach surpasses other spam-detection techniques.

9.1. The Proposed Framework vs. Roul et al. [13]. We compared the results of our framework with the work done by Roul et al. [13]. For the identification of spam web pages, they proposed a combined approach for content- and link-based spamdexing. To identify the content-based spamdexing, they used part of speech ratio test and term density, while for link-based spamdexing detection, they used the personalised page ranking to categorise the websites as spam and nonspam. For their experiments, they used the WEBSPAM-UK2006 dataset. Finally, they combined both of their spamdexing detection techniques to achieve 72.9% precision and 75.2% F-measure, listed in [47], which are significantly less than those found in our results.

9.2. The Proposed Framework vs. Dia et al. [47]. Next, we compared our empirical results with Dia et al. [47]. For spam identification, they considered the historical web page information in their work. For improvement in spam classification, they used content features from the old version of pages. They combined the classifiers based on the temporal characteristics and the current page content by applying supervised learning techniques. With their method, they extracted several temporal features from archival copies of the web presented by Internet Archive's Wayback Machine. For their experiments, they used the WEBSPAM-UK2007 dataset. Dai et al. achieved an F-measure of about 52.7 and a precision of approximately 65, which are less than what we achieved [47].

9.3. The Proposed Framework vs. Benczúr et al. [48]. Further, we compared the results obtained from our framework with the work of Benczúr et al. [48]. They introduced several features for web-spam filtering based on the appearance of keywords. The keywords they used were either highly advertised or frequently used in web spamming. Their web-spam features used for spamdexing detection include "the distribution of Google AdSense ads over pages of a site," "Google AdWords advertisement keywords suggestions," "the Yahoo Mindset classification of web pages," and "online commercial intention." To perform their experiments, they used the WEBSPAM-UK2006 dataset. They achieved an F-measure of 71.6% with a precision of 67.1, which are less than the results we obtained using our framework.

9.4. The Proposed Framework vs. Egele et al. [44]. Moreover, we compared our work with Egele et al. They introduced a technique to recognise the spam web pages on the search engine result pages. In the first step, the importance of different page features is determined by them to rank higher in the search engine result pages. Then, they developed a classification technique based on the page features to identify the spam web pages. For their experimental work, they used the J48 classifier. Finally, they listed the results of their experiments in Table 3 of Egele et al. [44]. They have achieved 51.2% precision and 41.9% F-measure, which are less than the results we obtained through our experiments.

9.5. The Proposed Framework vs. Becchetti et al. [49]. Finally, we compared our work with the results of Becchetti et al. [49]. Their technique used different content-based features, for instance, redirection to other pages, presence of unrelated keywords in URL, hidden text proportion on web pages, and duplicate content. Similarly, they plotted a web graph for link-based features and identified several page ranks like "trust rank", "degree-based measure," and "truncated page rank" of pages to identify the spam. As a base classifier, they have used the C4.5 decision tree. To perform their experiments, they used WEBSPAM-UK2002 and WEBSPAM-UK2006 datasets [49]. By combining both content- and link-based techniques, they achieved 66.8% precision and 72.3% F-measure, significantly less than those found in our results.

The comparisons above clearly show that our proposed framework for content- and link-based spamdexing detection is better than the above-mentioned five techniques.

10. Conclusion

Web spamming is a huge issue for people searching for information on the Internet using search engines. It also causes significant financial losses. Researchers have proposed many web-spam-detection methods to overcome this issue, but until now there is no single effective detection method that can detect all types of web spam available on the World Wide Web with high accuracy. In this research paper, we presented an improved combined approach for content- and link-based web-spam detection. We explored five different techniques for a content-based framework: stopword density, keyword density, spam keyword density, part of speech ratio, and copied-content test. Similarly, our link-based spamdexing detection framework used a paid-link database and spam-signal and link-farm identification technique with collaborative detection to detect a nonspam or spam page. We have used two datasets, WEBSPAM-UK2006 and WEBSPAM-UK2007, and the dataset obtained through our experimental work. An excellent and very promising F-measure of 79.6% compared to other existing approaches shows our framework's robustness. We will extend this research work by adding more content- and link-based spamdexing detection features to this framework. We believe that we can enhance our framework's power to identify a wide range of spam web pages by combining techniques and using more features.

Data Availability

WEBSPAM-UK2006 and WEBSPAM-UK2007 benchmark datasets are used in this paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the Office of Research, Innovation, Commercialization and Consultancy Management (ORICC), Universiti Tun Hussein Onn Malaysia (UTHM), and Ministry of Higher Education (MOHE), Malaysia, for financially supporting this Research under Tier-1 Research Grant vote no. H938.

References

[1] Z. Gyongyi and H. Garcia-Molina, "Web spam taxonomy," in *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005)*, Chiba, Japan, 2005.

[2] M. R. Henzinger, R. Motwani, and C. Silverstein, "Challenges in web search engines," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003.

[3] 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*, p. 83, Scotland, UK, 2006.

[4] N. Z. J. MCA and P. Prakash, "Document content based web spam detection using cosine similarity," *International Journal of Intelligence Research (IJOIR)*, vol. 7, 2016.

[5] A. Shahzad, N. Mohd Nawi, E. Sutoyo et al., "Search engine optimization techniques for malaysian university websites: a comparative analysis on google and bing search engine," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 8, no. 4, pp. 1262–1269, 2018.

[6] 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.

[7] Z. Guo and Y. Guan, "Active probing-based schemes and data analytics for investigating malicious fast-flux web-cloaking based domains," in *Proceedings of the 2018 27th International Conference on Computer Communication and Networks (ICCCN)*, pp. 1–9, Hangzhou, China, August 2018.

[8] N. Spirin and J. Han, "Survey on web spam detection," *ACM SIGKDD Explorations Newsletter*, vol. 13, no. 2, pp. 50–64, 2012.

[9] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, pp. 42–49, Washington, DC, USA, November 2004.

[10] C. Zhai, "Statistical language models for information retrieval," *Synthesis Lectures on Human Language Technologies*, vol. 1, no. 1, pp. 1–141, 2008.

[11] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[12] N. El-Mawass and S. Alaboodi, "Data quality challenges in social spam research," *Journal of Data and Information Quality*, vol. 9, no. 1, pp. 1–4, 2017.

[13] R. K. Roul, S. R. Asthana, M. Shah, and D. Parikh, "Detecting spam web pages using content and link-based techniques," *Sadhana*, vol. 41, no. 2, pp. 193–202, 2016.

[14] B. Davison, "Recognising nepotistic links on the web," *Artificial Intelligence Web Search*, pp. 23–28, 2000.

[15] J. Piskorski, M. Sydow, and D. Weiss, "Exploring linguistic features for web spam detection: a preliminary study," in *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, pp. 25–28, Beijing, China, April 2008.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[17] Y. Tian, G. M. Weiss, and Q. Ma, "A semi-supervised approach for web spam detection using combinatorial feature-fusion," in *Proceedings of the Graph Labelling Workshop and Web Spam Challenge at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery*, pp. 16–23, New York, NY, USA, September 2007.

[18] D. Fetterly, M. Manasse, and M. Najork, "Detecting phrase-level duplication on the world wide web," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 170–177, Salvador, Brazil, August 2005.

[19] D. Fetterly, M. Manasse, and M. Najork, "On the evolution of clusters of near-duplicate web pages," in *Proceedings of the IEEE/LEOS 3rd International Conference on Numerical Simulation of Semiconductor Optoelectronic Devices (IEEE Cat. No. 03EX726)*, pp. 37–45, IEEE. Tokyo, Japan, 2003.

[20] 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: colocated with ACM SIGMOD/PODS 2004*, pp. 1–6, Paris, France, June 2004.

[21] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," *Computer Networks ISDN System*, vol. 29, no. 8, pp. 1157–1166, 1997.

[22] A. Z. Broder, "Some Applications of Rabin's Fingerprinting Method," in *Sequences II*, pp. 143–152, Springer, New York, NY, USA, 1993.

[23] M. O. Rabin, "Fingerprinting by random polynomials," *Center for Research in Computing Technology*, Report TR-15-81, Harvard University, Cambridge, MA, USA, 1981.

[24] M. Erdélyi, A. Garzó, and A. A. Benczúr, "Web spam classification: a few features worth more," in *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*, pp. 27–34, Hyderabad, India, 2011.

[25] T. Urvoy, T. Lavergne, and P. Filoche, "Tracking web spam with hidden style similarity," in *Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web*, pp. 25–31, Seattle, WA, USA, August 2006.

[26] D. Hiemstra, *Language Models BT—Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds., Springer US, Boston, MA, USA, 2009.

[27] G. Mishne, D. Carmel, and R. Lempel, "Blocking blog spam with language model disagreement," *AIRWeb*, vol. 5, pp. 1–6, 2005.

[28] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang, "Knowing your enemy: understanding and detecting malicious web advertising," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, pp. 674–686, Raleigh, NC, USA, October 2012.

[29] B. Stone-Gross, R. Stevens, A. Zarras, R. Kemmerer, C. Kruegel, and G. Vigna, "Understanding fraudulent activities in online ad exchanges," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pp. 279–294, Berlin, Germany, November 2011.

[30] A. Zarras, A. Kapravelos, G. Stringhini, T. Holz, C. Kruegel, and G. Vigna, "The dark alleys of madison avenue: understanding malicious advertisements," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, pp. 373–380, Vancouver, Canada, November 2014.

[31] S. Antonatos, I. Polakis, T. Petsas, and E. P. Markatos, "A systematic characterization of IM threats using honeypots," in *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, February 2010.

[32] J. L. Ledford, *Search Engine Optimisation Bible*, John Wiley & Sons, Hoboken, NY, USA, 2015.

[33] P. T. Metaxas and J. DeStefano, "Web spam, propaganda and trust," in *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, Chiba, Japan, May 2005.

[34] S. Sethi and A. Dixit, "A novel page ranking mechanism based on user browsing patterns," *Software Engineering*, Springer, Berlin, Germany, pp. 37–49, 2019.

[35] D. Sharma, R. Shukla, A. K. Giri, and S. Kumar, "A brief review on search engine optimisation," in *Proceedings of the 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 687–692, Noida, India, 2019.

[36] M. D. 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.

[37] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 194–201, 2017.

[38] S. Nomura, S. Oyama, T. Hayamizu, and T. Ishida, "Analysis and improvement of hits algorithm for detecting web communities," *Systems and Computers in Japan*, vol. 35, no. 13, pp. 32–42, 2004.

[39] J. J. Whang, Y. S. Jeong, I. S. Dhillon, S. Kang, and J. Lee, "Fast asynchronous anti-trust rank for web spam detection," in *Proceedings of the WSDM Worskshop MIS2*, Marina Del Rey, CA, USA, 2018.

[40] J. Abernethy, O. Chapelle, and C. Castillo, "Graph regularisation methods for web spam detection," *Machine Learning*, vol. 81, no. 2, pp. 207–225, 2010.

[41] Q. Gan and T. Suel, "Improving web spam classifiers using link structure," in *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, pp. 17–20, Banff, Canada, May 2007.

[42] V. M. Prieto, M. Álvarez, and F. Cacheda, "SAAD, a content based web spam analyser and detector," *Journal of Systems and Software*, vol. 86, no. 11, pp. 2906–2918, 2013.

[43] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer, "The connectivity sonar: detecting site functionality by structural patterns," in *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, pp. 38–47, Nottingham, UK, August 2003.

[44] M. Egele, C. Kolbitsch, and C. Platzer, "Removing web spam links from search engine results," *Journal in Computer Virology*, vol. 7, no. 1, pp. 51–62, 2011.

[45] K. L. Goh, R. K. Patchmuthu, and A. K. Singh, "Link-based web spam detection using weight properties," *Journal of Intelligent Information Systems*, vol. 43, no. 1, pp. 129–145, 2014.

[46] A. Shahzad, H. Mahdin, and N. M. Nawi, "An improved framework for content-based spamdexing detection," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 11, no. 1, 2020.

[47] X. Dai, B. D. Davison, and X. Qi, "Looking into the past to better classify web spam," in *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, pp. 1–8, Madrid, Spain, April 2009.

[48] A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós, "Web spam detection via commercial intent analysis," in *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, pp. 89–92, Banff, Canada, May 2007.

[49] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, "Web spam detection: link-based and content-based techniques," *European Integral Project Dynamic Evolution Large Scale Information System DELIS Proceeding Final Work.*vol. 222, pp. 99–113, 2008, https://www.chato.cl/papers/becchetti_2008_link_spam_techniques.pdf.