Network and Complex Systems
ISSN 2224-610X (Paper) ISSN 2225-0603 (Online) DOI: 10.7176/NCS
Vol.10, 2019

Performance Evaluation of User-Behaviour Techniques of Web Spam Detection Models

Oluwatoyin Odukoya, Bodunde Akinyemi*, Mohammed Fofana, Ganiyu Aderounmu
Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria
* E-mail of the corresponding author: bakinyemi@oauife.edu.ng

Abstract

Web spam detection is a critical issue in today's rapidly growing usage of the Internet and the World Wide Web. The upsurge of web spam has significantly deteriorated the Quality of Services (QoS) of the World Wide Web. The degeneration of the quality of search engine results has given rise to researches on the detection of spam pages efficiently and accurately. Existing user-behaviour oriented web spam detection models employed the content-based, link-based and other features of webpages for classification of web spams. These user-behaviour techniques either implemented singly or combined has achieved good detection performance. However, the effectiveness of these features in identifying Web spams correctly needs to be determined. In this study, predictive web spam detection models that employed all related user-behaviour features of webpages were developed and evaluated. The content, link, and obvious-based features datasets were collected from an online repository. Relevant features were extracted using an improved Filter-based method. Six user-behaviour related features extracted from the datasets were used to combine the datasets to generate all possible subset of feature space required, such that 7 new datasets were generated for the study. Multi-Layer Perceptron (MLP) approach was adopted as a classifier for each of the identified features. Python Machine Learning Library was used to simulate the models using percentage splits of 60/40%, 70/30% and 80/20% ratio for training/testing dataset and the performances were evaluated using accuracy, True Positive (TP) rate, False Positive (FP) rate and precision as metrics. The result showed that for the majority of the datasets the formulated models have shown an increase in efficiency after feature selection. The MLP classifier was able to achieve the best result of 66.0% accuracy when the link-based dataset was used with feature selection. The study concluded that link-based features of a user is sufficient and effective for the detection of web spams.

Keywords: Webspam, Content-based, Link-based, features, user-behaviour, evaluation

DOI: 10.7176/NCS/10-07
Publication date:December 31st 2019

1. Introduction

Web Spams are unsolicited, unwanted email, ads, links, contents, sent indiscriminately, directly or indirectly by a sender having no current relationship with the recipient, or an unsolicited commercial mail usually sent to a large group of recipients at the same time by service providers such as Internet Service providers (ISPs) (Ndumiyana *et al*., 2013). Web spams are usually popped up during the search for information on the web, these ads or junks are developed by spammers to attract web users and cause a search engine to produce wrong information when surfing web or websites (Jindal and Liu, 2008). Some of the nefarious act posed by web spam includes subverting the ranking algorithms of web search engines and cause them to rank search results higher than they would otherwise (Najork, 2009) etc.

This spam situation is so disruptive and infuriating that search engines, web users, and email receivers spend a lot of time trying to combat it since it leads to the loss of lots of resources, finance, and cost. commercial search engines treat their precise set of spam-prediction features as extremely proprietary, and features (as well as spamming techniques) evolve continuously as search engines and web spammers are engaged in a continuing arms race (Manne and Wright, 2011).

There have been different classification problems amenable to machine learning techniques to combat this deadly web masquerade. Spam classifiers and filterers are been created in search to combat web spam, but due to the smartness of the spammers, they always develop new ways to manipulate their way into search engines (Castillo *et al.,* 2007). Spam classifiers and filterers take a large set of diverse features as input, including content-based features, link-based features, DNS and domain-registration features, and implicit user feedback. However, it was observed that the majority of the classifiers focused on content and/or link-based features of the

webpages for detecting web spam. Many a time, most of these algorithms are not pro-active and cannot withstand the pressure from spam pages, because if the spams links, emails, web pages are blocked by the filterer's or classifiers for the first time, within no time the spam will replicate itself and hit the algorithm for as many times until it makes its way through.

Meanwhile, Web users' interests, navigational actions, and preferences have gained importance in web spam detection, since web users contribute greatly to the sharing of spam pages, spam sites and also making spam pages or sites to gain more relevance. Most users who visit the web are not abreast of spam pages or sites, because of these reasons their navigate Pattern is always a problem by clicking on every page that come up with interesting topics. The navigational patterns of a user are stored within web access logs, and these contain many different data in these files. In order to understand the user behaviour through these data that are stored in the web access logs, some of the existing studies reviewed the extraction of user behaviour data from web-pages and web access log files but adopted the use of expert knowledge for the classification of web spam, while some employed web usage mining concept, which consists of the application of machine learning techniques over data originated in the Web (Web data) for automatic extraction of behavioural patterns from Web users (Román *et al.,* 2014).

Implementation of a web spam detection model using user behaviour-related data has greatly helped improve the prevention of illegal and unsolicited access to web spam pages from a website that is a spam (Liu *et al*. 2008). The availability of web access log files provided a means of understanding the interaction of users with web pages to ascertain information about the status of webpages visited. Concerning user behavior analysis, there have been several challenges in combating web spam in search engines and online social networks. The challenge of detecting newly appearing web spams called Zero-day spams. For this reason, many times the users of the World-Wide-Web still encounter embarrassment from spam pages, adds, pop up, links redirections, false news, porn site, etc. In the end, most people might not likely get what they are searching for from the web, due to the deceitful natures of the spamming community. Thus, there is a need to determine which of the user behaviour techniques is effective for detecting newly appearing web spams.

In this study, an attempt was made to evaluate the effectiveness and robustness of a web spam detection model that employed the existing content, link, and obvious-based features complemented with other user-behaviour features. The model with the best performance will be selected for the detection of web spam.

The rest of the paper is organized as follows: related works were discussed in section2 while section3 discusses the methodology used to solve the identified problem. Section 4 discusses the simulation process and results. Section 5 offers conclusions and recommendations for future work.

2. Related Works

Research on web spam detection has been ongoing for over a decade. Several web spam detection algorithms have been developed to identify different type of spam that appears on the Web. Spirin and Han (2011), Castillo and Davison (2011), Kohle and Bhukte (2015), gave a comprehensive survey on the state of the art of the different techniques used for web spam detection. With respects to user behavior analysis, web spams are detected either by analysing the content-based features of the web page contents. |Some studies adopted the use of the content-based features to detect web spams (Fetterly *et al.*, 2004; Gyongyi and Garia-Molina, 2005; Mishne, 2005; Ntoulas *et al*., 2006; Svore, 2007; Sydow, 2007; Liu *et al.,* 2008; Piskorski, 2008; Erdelyl *et al*., 2009; Awad and Elseoufi, 2011; Erd´elyi, 2011; Iqbal and Abid, 2015; Rao *et al*., 2016; Al-Zoubi *et al.,* 2017). Some studies adopted the use of hyperlink structure analysis (Davison, 2000; Amitay *et al.,* 2003; Caverlee and Liu, 2007; Bencz´ ur *et al*., 2006; Baeza-Yates *et al*., 2006; Becchetti *et al.,* 2006; Niu *et al.,* 2018; Hochbaum *et al*., 2019). Some studies adopted the use of combinations of content-based and link-based features for Web Spam detection (Gyongyi *et al.,* 2004; Wu and Davison, 2006; Castillo *et al*., 2007; Liu *et al*., 2011). Whereas, some study adopted the use of other features like click-based and posting-based features for Spam URL Detection (Wei *et al* 2012; Cao and Caverlee, 2015)

All these user-behaviour oriented algorithms showed a competitive performance in detection, maximize the accuracy and minimize the false positives rates, which demonstrates the successful results of many researchers. Despite all these successes, spam is still constantly evolving and affects many people and businesses still negatively. Thus, the need to evaluate the effectiveness of the user behaviours features to identify which one works best in detecting zero-day spams.

3. Methodology

Figure 1 shows the conceptual diagram of the study and the methods adopted to achieve the objectives of the study are as follows:

(a) Collection of data from an online repository.

(b) Identification and analysis of the web-usage features and the user-behaviour features required for assessing web-pages.

(c) Formulation of the spam detection model for each of the identified features in the dataset.

(d) Simulation of the models developed from these datasets using percentage splits of 60/40 and 70/30, 80/20 percentage for training/testing set selected from data collected.

(e) Performance evaluation of the models using accuracy, true positive (TP) rate, false positive (FP) rate and precision to select the most appropriate user behavior technique for web spam detection.

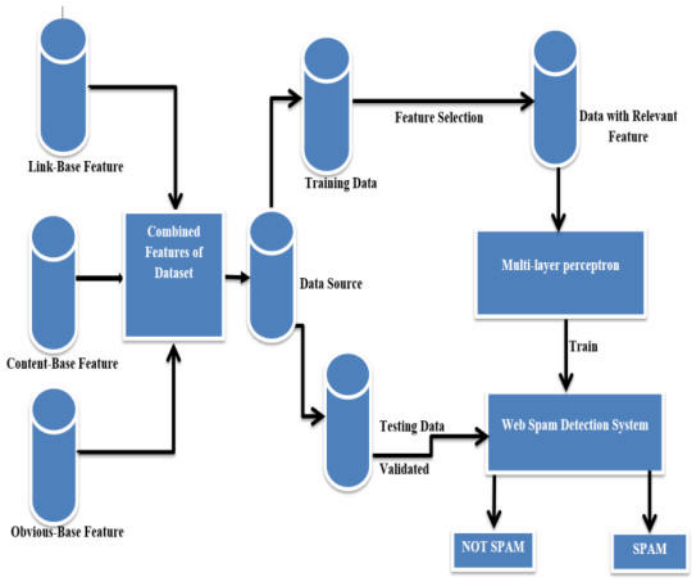

Figure 1: Conceptual Diagram of the Proposed Model

3.1 Data Collection and Analysis

A web spam dataset called the UK-2006-Web-Challenge Data was collected from the Web Spam Challenge website. It contained 3 classes of the web-usage feature-based dataset, namely: content-based features, link-based features and obvious features. The distribution of the target class used to describe each host that was assessed are as shown in Table 1, with Content-based features.csv.gz, Link-based-features.csv.gz and Obvious-features.csv.gz consisting of 3849, 3998 and 3849 host records respectively.

Table 1: Distribution of Target Class among assessed Web Hosts

Dataset Class	Spam Host	Non-Spam Host	Total
Content-Based Features	220	3629	3849
Link-Based Features	222	3776	3998
Obvious Features	220	3629	3849

3.2 Features Extraction Process

Filter-based feature selection (FS) methods were employed to determine the relevant features from the datasets, FS methods define relevance by identifying the attributes that are more correlated with the target class (spam or non-spam) and they are also less computationally expensive. A backward elimination technique which began by selecting all the initially identified feature set in the dataset and evaluates their accuracy using a classifier was applied. The process is depicted in Figure 2. This process was repeated for every possible subset of features in the dataset by progressive elimination of features until an empty space of features. The feature set evaluated with the highest accuracy is returned as the most relevant set of features that are required to improve the performance of the proposed spam detection model.

The dataset containing the 3 web-usage features were combined to generate seven (7) possible permutations of the datasets, the features of the datasets were used as a basis for combining the dataset to generate all possible subset of feature space required for the model development. The datasets were combined such that 3 datasets consisted of sets of content-based only, obvious only and link-based only, 3 dataset consisted of sets of content and link-based, content and obvious, and obvious and link-based while 1 dataset consisted of the set of content-based, link-based and obvious features as shown in Table 2 making a total of 7 datasets adopted for this study. Figure 3 shows a description of the seven (7) various feature class-based dataset generated from the three (3) dataset collected.

Also, the user-behaviour features (proposed by Liu *et al*., 2012) were extracted from the initial features collected from the web-usage features. They are the six user-behaviour features that can be adopted for separating spam pages from ordinary ones. The first five are from user-behaviour patterns, and the last feature is a link-analysis feature extracted from a user-browsing graph that is also constructed with users' Web access log data. They are described as follows:

a. Search Engine Oriented Visit Rate (SEOV) – The number of user visits that are oriented from search engine.

$$SEOV(p) = \frac{(no \text{ of search engine oriented visits of webpagep})}{(no \text{ of visits of webpage p})} \quad (1)$$

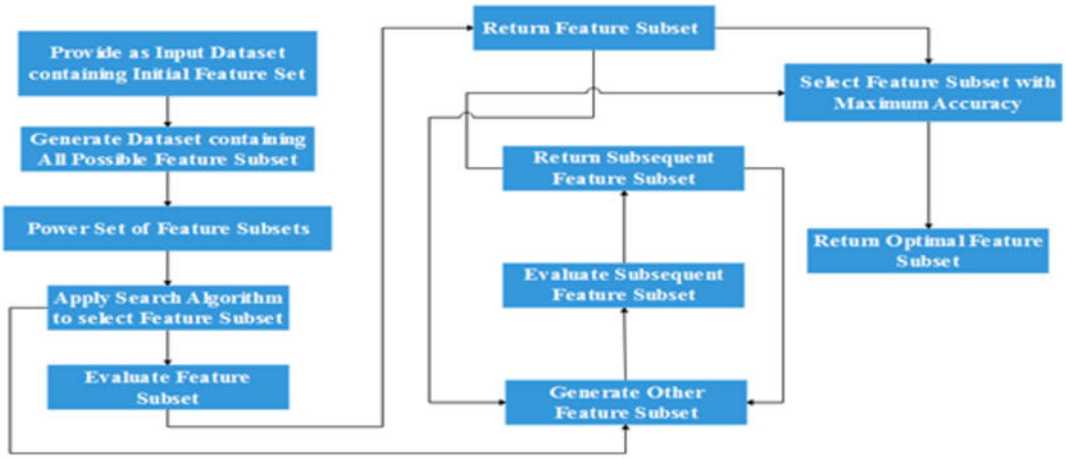

Figure 2: Feature Selection Process

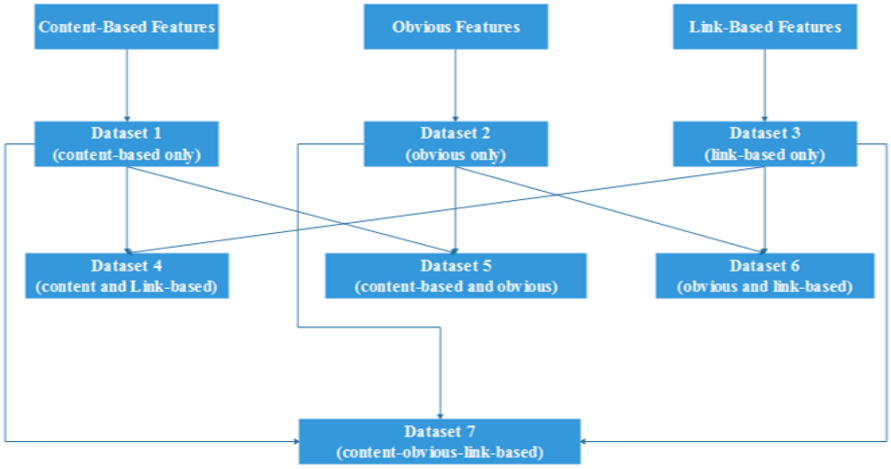

Figure 3: Features Combination Process

b. Source Page Rate (SPG) – The number of users that follow links on the page

$$SP(p) = \frac{(no \text{ web page p that appears as the source page})}{(\text{no of webpage p appears in the Web access logs})} \quad (2)$$

c. Short-Time Navigation Rate (SN) – The number of users that will not visit the site in the future

$$SN(p) = \frac{(no \text{ of sessions in which users visit less than N pages in p})}{(\text{no of sessions in which users visit webpage p})} \quad (3)$$

d. Query Diversity – The number of user visits that are oriented by hot keyword searches

$$QD(p) = No \text{ of query topics that lead user visit for webpage p} \quad (4)$$

e. Spam Query Number (SQN) – The number of pages that a certain user visit in the site

$$SQN(p) = No \text{ of spam query terms that lead to user visit for p} \quad (5)$$

f. User-Oriented TrustRank – The number of users that visit the site

Table 2 shows the description of the classification system that was used to identify each user behaviour feature adopted for Web Spam classification.

Table 2: Classification of Web Pages based on User behaviour Features

User Behaviour Feature	Spam Websites	Non-Spam Websites
Search Engine Oriented Visit Rate (SEOV)	High	Low
Source Page Rate (SPG)	Low	High
Short-Time Navigation Rate (SN)	Low	High
Query Diversity	High	Low
Spam Query Number (SQN)	High	Low
User-Oriented TrustRank	Low	High

3.3 Model Formulation

The predictive model adopted for this study was the Multi-Layer Perceptron (MLP) as described by Idowu *et al*., (2019). The MLP is composed of three (3) main layers, namely: the input, hidden and output layer. The MLP consisted of *n* input layers which were proportional in value to the number of features identified in the dataset presented. The hidden layer consisted of *a* layers of which consisted of neuron which received input from the input layers and produced outputs via an activation function which was used to produce outputs and propagated to another neuron in subsequent layers. The output layer consisted of two (2) neurons which represented the target class for identifying spam and non-spam web hosts from the dataset.

The formulated predictive model for web spam detection using the features selected is presented in Figure 4. A mapping function was used to express the process of model formulation from the feature space to the output space. The training dataset *S* which consisted of the initial features identified at the point of data identification and collection is represented by X_i, where i is the number of features existing in the original dataset of web hosts, and X'_j consists of the features relevant for predicting web spam, such that: $j < i$. The process of feature selection is represented by the mapping function, F in equation (6).

$$F: X_i \rightarrow X'_j \tag{6}$$

Such that: X_i are the original set of attributes collected and X'_j are the relevant features selected by the feature selection method. Following the process of feature selection, the new dataset belongs X^i_{jk} such that k is the number of web hosts records collected in the original dataset. If *n* datasets were selected for training the predictive model using a supervised machine learning to formulate the model using the relevant variables using the mapping in equation (7).

$$\varphi: X_{jk} \rightarrow Y_k;$$
$$defined\ as\ \varphi(X_{jk}) = Y_k\ for\ all\ web\ host\ records, k \tag{7}$$


```
   Input: Data
   Output: Spam, Ham
 1 initialization;
   /* read data fro the directory                                    */
 2 dataset = 'datasets/Link-features-only-new.csv';
   /* function to process data                                       */
 3 def get_processdata(data):
 4 |   dataset = pd.read_csv(data);
   |   /* rename class variables, non-spam to spam                  */
 5 |   X = dataset.iloc[:, :-1];
 6 |   y = dataset.ilco[:, -1];
 7 |   return X, y ;
   |   /* perform data normalization on the X data                  */
   |   /* divide data into training and setests                     */
 8 |   Xtrain, Xtest, ytrain, ytest = train_test_split(X.values, y.values,
   |    test_size=(tss/100), random_state=7);
   |   /* fit the training set into the algorithm and predict
   |      the testset                                               */
   /* perform testing for unknown data                               */
 9 def predict_spam(test):
10 |   prediction, accuracy, model = fit_predict(trainset, testset, ytest,
   |    scaler);
11 |   predict = model.predict(test.values);
12 |   result = [];
13 |   for p in predict do
14 |   |   if p == 0 then
15 |   |   |   result.append('Ham');
16 |   |   else
17 |   |   |   result.append('Spam');
18 |   |   end
19 |   end
20 |   return result;
   /* evaluate the system with the machine learning metrics
      */
   /* evaluate the system with the confussion metrics                */
   /* plot the roc curve                                             */
```

Figure 4: Proposed Web Spam Detection Model

Such that: X_{jk} represented the set of attributes, j for web hosts record, k and Y_k is the target class (spam or non-spam) of record, k. Therefore, the supervised machine learning algorithm considered in this study was expected to determine the best fit for $\varphi \in \mathbb{H}$ (the set of all possible models) based on the minimization of the cost function defined according to equation (8).

$$\mathbb{L}: \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{Z};$$

$$defined\ as\ \mathbb{L}(Y_{a,}Y_p) \tag{8}$$

Such that: $\mathbb{Z}$ is a set $[0,1]$ and $Y_{a,}Y_p$ are the actual and predicted values of the target class respectively. Since, the problem is a classification problem that would involve the identification of a value the value of 0 implied correct while the value of 1 implied incorrect classification. Hence, the ability to classify correctly the detection of web spam is determined according to the cost function defined in equation (9).

$$\mathbb{L}(Y_{a,}Y_p) = \begin{cases} correct\ classification; & = 0 \\ incorrect\ classifcation; & = 1 \end{cases} \tag{9}$$

4. Results and Discussion

The detailed results of this study are as follows:

4.1 Data Analysis results

The data collected contained 3999 records of web pages which were assessed as spam and non-spam pages based on the user-behaviour scores alongside the features identified from the 3 classes of datasets collected. The descriptions of the feature Sets of the 3 Datasets are treated offline in this paper. The dataset was used as a basis of extracting six (6) user behaviour related features for the classification of each web host as either Spam or non-Spam. This approach was based on the principle of the *Wisdom of the Crowds* which focused on using the interaction of users with the webpages to determine the nature of the webpages visited based on the characteristic nature of the six (6) user-behaviour features identified. The results of the classification of the three (3) dataset collected for this study are presented in Table 3.

4.2 Features Selection Result

The process of feature selection was repeated for all the 7 set of the dataset with the most relevant feature set for each dataset identified as shown in Table 4. Table 5 shows the presence of the number of features in each dataset alongside the number and the proportion of features selected. The results showed that 49.5%, 50.7%, 33.3%, 49.6%, 49.5%, 49.3%, 49.6% features were selected for content-based, link-based, obvious-based, content and link-based, content and obvious-based, link and obvious-based, Content, Link and Obvious datasets respectively.

Table 3: Distribution of the Dataset Identified and Collected for Web Spam Detection

Dataset	Number of Features	Non-Spam		Spam	
		Frequency	Percentage	Frequency	Percentage
Link Features Only	138	250	53.19	220	46.81
Content Features Only	97	250	53.19	220	46.81
Obvious Features Only	3	250	53.19	220	46.81

Table 4: Distribution of Dataset containing a combination of Features

Dataset Name	Feature Name Tag	Number of Features
Content-Based	C	97
Link-Based	L	138
Obvious	O	3
Content and Link-Based	CL	231
Content and Obvious	CO	97
Link and Obvious	LO	137
Content, Link and Obvious	CLO	233

Table 5: Results of the selection of relevant Features from Datasets

Dataset Name	Number of Features	Number of Selected Features	The proportion of Features (%)
Content Based	96	48	49.48
Link-Based	137	70	50.72
Obvious	2	1	33.33
Content and Link-Based	231	116	49.57
Content and Obvious	97	49	49.49
Link and Obvious	137	69	49.29
Content, Link and Obvious	233	117	49.58

4.3 Model Simulation and Evaluation Result

The proposed model was simulated using the Python Machine Learning Library. The Seven (7) datasets were split into training and testing datasets according to a proportion of 60%/40%, 70%/30%, and 80%/20%. Table 6 shows the distribution of the target classes (spam and non-spam) within each proportion of the training and testing dataset collected for this study. Using the full features set and the selected features of the 7 datasets, a total of 14 datasets were subjected to a training and testing scheme of 60/40, 70/30 and 80/20 process. Thus, 42 simulations were carried out such that there were 14 simulations for each training scheme. The results of the simulations are presented for each training and testing scheme as follows:

(i.) Using 60/40 percentage scheme

Table 7 and Figure 5 presents the results of the simulation of the proposed classification model using the 60/40 percentage split scheme performed on the 7 datasets using the full feature set and relevant selected feature set containing 188 records in the testing dataset. It was shown that the MLP classification model for web spam detection had the best classification result using the link-based dataset with the relevant features selected using the feature selection process.

(ii.) Using a 70/30 percentage scheme

Table 8 and Figure 6 presents the results of the simulation of the proposed classification model using the 70/30 percent split scheme performed on the 7 datasets using the full feature set and relevant selected feature set containing 141 records in the testing dataset. It was shown that the MLP classification model for web spam detection had the best classification result using the link-based dataset with the originally identified features (without FS).

(iii.) Using the 80/20 percentage scheme

Table 9 and Figure 7 presents the results of the simulation of the Proposed classification model using the 80/20 percent split performed on the 7 datasets using the full feature set and relevant selected feature set containing 94 records in the testing dataset. It was shown that the MLP classification model for web spam detection had the best classification result using the link and obvious-based dataset with the originally identified features (without FS).

From the evaluation results, it was shown that the model with the best overall performance was identified as the multi-layer perceptron (MLP) classifier that was modelled using the dataset containing the initially identified link-based features alone.

Table 6: Distribution of the Dataset for Training and Simulation

Class Label	Simulation I		Simulation II		Simulation III		Total
	Training	Testing	Training	Testing	Training	Testing	
Spam	126	94	154	66	176	44	220
Non-Spam	126	94	175	75	200	50	250

Table 7: Simulation results using 60/40 Percentage Scheme

Dataset	Feature Selection process	Total Correct	Accuracy %	TP rate		FP rate	
				Ham	Spam	Ham	Spam
Content	Without FS	96	51.06	0.134	0.912	0.088	0.866
	With FS	88	46.81	0.423	0.516	0.484	0.577
Content-Link	Without FS	101	53.72	0.474	0.604	0.396	0.526
	With FS	112	59.57	0.691	0.495	0.505	0.309
Content-Link-Obvious	Without FS	95	50.53	0.856	0.132	0.868	0.144
	With FS	115	61.17	0.711	0.505	0.495	0.289
Content-Obvious	Without FS	98	52.13	0.557	0.484	0.516	0.443
	With FS	91	48.4	0.464	0.505	0.495	0.536
Link	Without FS	119	63.3	0.526	0.747	0.253	0.474
	With FS	**124**	**65.96**	**0.732**	**0.582**	**0.418**	**0.268**
Link-Obvious	Without FS	107	56.91	0.381	0.769	0.231	0.619
	With FS	117	62.23	0.722	0.516	0.484	0.278
Obvious	Without FS	94	50	1	0	1	0
	With FS	115	61.17	0.872	0.351	0.649	0.128

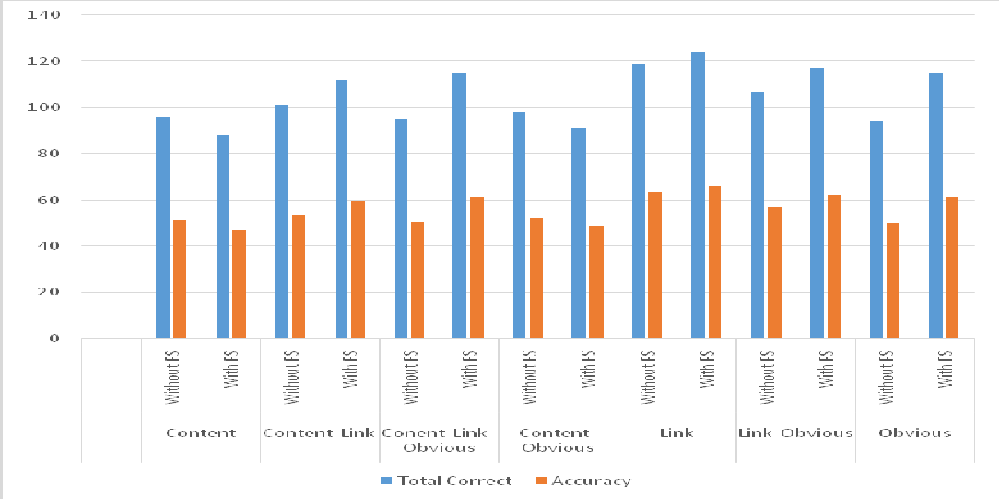

Figure 5: Classifications and Accuracy results using 60/40 Percentage Split

Table 8: Simulation results using 70/30 Percentage Scheme

Dataset	Feature Selection process	Total Correct	Accuracy %	TP rate		FP rate	
				Ham	Spam	Ham	Spam
Content	Without FS	74	52.48	0.157	0.887	0.113	0.843
	With FS	69	48.94	0.357	0.62	0.38	0.643
Content-Link	Without FS	71	50.35	0.371	0.634	0.366	0.629
	With FS	73	51.77	0.543	0.493	0.507	0.457
Content-Link-Obvious	Without FS	61	56.74	0.486	0.648	0.352	0.514
	With FS	73	51.77	0.557	0.479	0.521	0.443
Content-Obvious	Without FS	85	60.28	0.6	0.606	0.394	0.4
	With FS	62	43.97	0.457	0.423	0.577	0.543
Link	Without FS	97	68.79	0.771	0.606	0.394	0.229
	With FS	91	64.54	0.729	0.563	0.437	0.271
Link-Obvious	Without FS	72	51.06	0.729	0.296	0.704	0.271
	With FS	89	63.12	0.743	0.521	0.479	0.257
Obvious	Without FS	79	56.03	0.303	0.787	0.213	0.697
	With FS	78	55.32	0.318	0.76	0.24	0.682

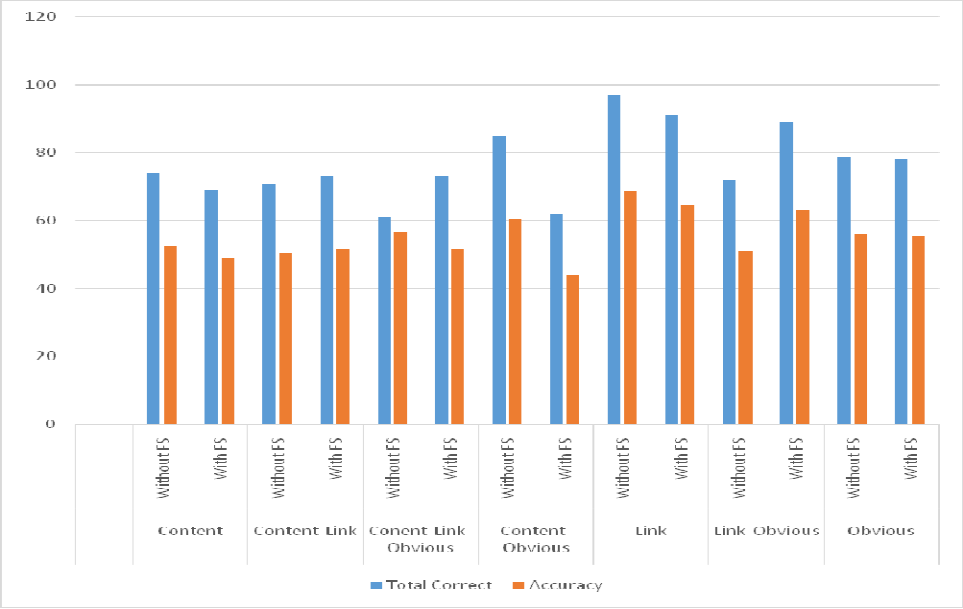

Figure 6: Classifications and Accuracy result using 70/30 Percentage Split

Table 9: Simulation results using 80/20 Percentage Scheme

Dataset	Feature Selection Process	Total Correct	Accuracy %	TP rate		FP rate	
				Ham	Spam	Ham	Spam
Content	Without FS	53	56.38	0.048	0.981	0.019	0.952
	With FS	47	50	0.429	0.558	0.442	0.571
Content-Link	Without FS	49	52.13	0.667	0.404	0.596	0.333
	With FS	52	55.32	0.643	0.481	0.519	0.357
Content-Link-Obvious	Without FS	51	54.26	0.81	0.327	0.673	0.19
	With FS	53	56.38	0.69	0.462	0.538	0.31
Content-Obvious	Without FS	42	44.68	1	0	1	0
	With FS	46	48.94	0.429	0.538	0.462	0.571
Link	Without FS	57	60.64	0.667	0.558	0.442	0.333
	With FS	58	61.7	0.738	0.519	0.481	0.262
Link-Obvious	Without FS	60	63.83	0.619	0.654	0.346	0.381
	With FS	57	60.64	0.714	0.519	0.481	0.286
Obvious	Without FS	55	58.51	0.275	0.815	0.185	0.725
	With FS	55	58.51	0.175	0.889	0.111	0.825

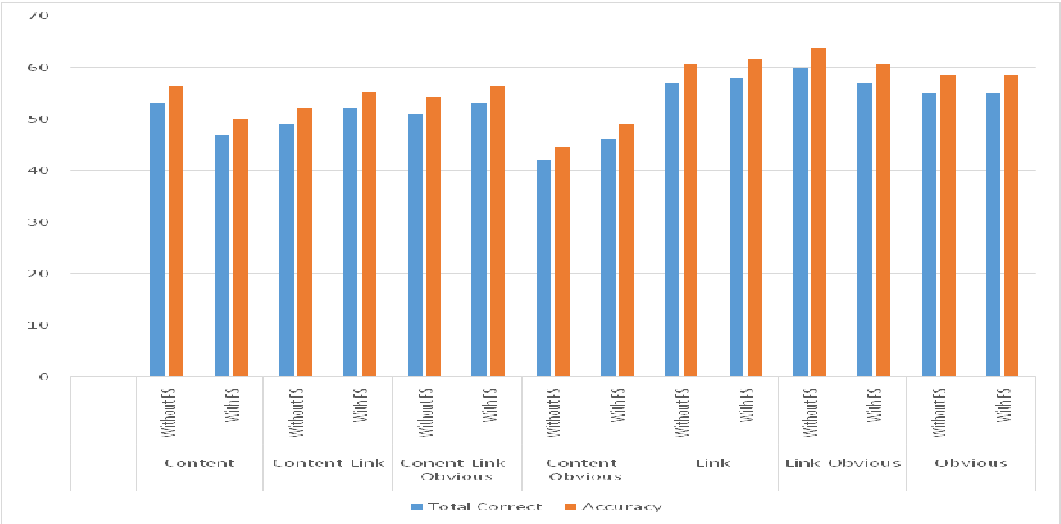

Figure 7: Classifications and Accuracy Result using 80/20 Percentage Split

Figure 8 shows the results of the TP rate, FP rate and precision of the spam (ham) and spam web-hosts for the MLP classifiers with the best performance for each percentage-split training scheme. The results showed that the MLP classifier using the 70/30 percentage split scheme (without FS) had the highest TP rate for non-spam web-hosts and moderate for spam web-hosts. The results also showed that the MLP classifier using the 70/30 percentage split (without FS) had a moderate FP rate for non-spam web-hosts and lowest for spam web-hosts.

In summary, the results of the study showed that the feature selection algorithm adopted for this study selected 50% of the initially identified features as relevant. The simulation results showed that the MLP classification model for web spam detection had the best classification perfomance using the link-based dataset with the relevant features selected using the 60/40 and 80/20 percentage split, and using the link-obvious based dataset with the originally identified features for the 70/30 percentage split. The result showed the accuracy of the MLP with and without features selection using the link and obvious-based dataset showed an accuracy of 63.8% however using content and obvious-based dataset showed an accuracy of 48.9%, and content and link-based dataset with the accuracy of 55.3%. Using the three combined features consisting of content, link and obvious-based datasets with and without features selection showed an accuracy of 56.4%. Using the isolated features dataset for simulation by the MLP with and without feature selection, the content-based dataset showed an accuracy of 56.4%, link-based dataset showed an accuracy of 61.7% which was better than content-based dataset while the obvious-based dataset showed an accuracy of 58.5% which also outperformed the content-based dataset.

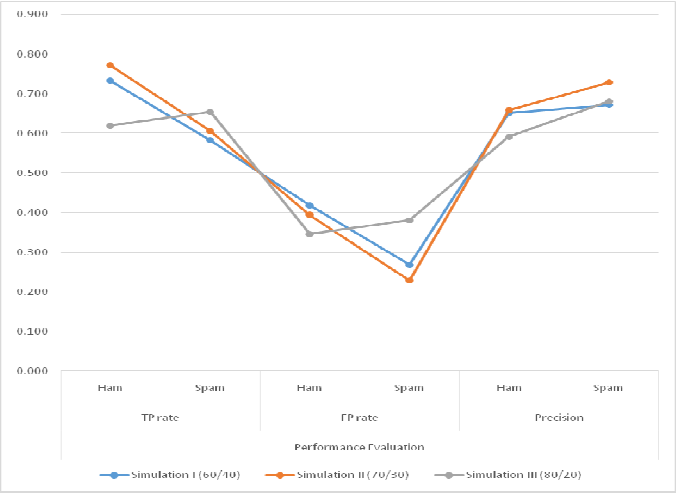

Figure 8: Evaluation results of the Models

5. Conclusion

This study evaluated the accuracy performances of different user-behaviour features used in modelling web spam detection models. The study identified the user-behaviour features from a selected dataset and then developed a classification model for the detection of spam websites using Multi-Layer Perceptron (MLP). The classification models were simulated and validated. The results showed that the MLP classifier using the 70/30 percentage split scheme (without FS) had the highest precision for non-spam web-hosts and spam web-hosts. The study proved that the Link-based features integrated into a classification model will facilitate more effective detection of spam websites as compared to content-based features. This is justified since content-based features only contain details about the content of the webpages, however, the link-based features reveal information about the connection of pages alongside the navigation of users through these paths. Therefore, users are likely to spend lesser time on a website which also implies lesser movement through the path of websites created by the hyperlinks. This will ensure that unsuspecting users are not directed to the contents of such websites thus mitigating the risk associated with visiting such spam websites.

The timeliness problem is still a challenge. Suggested future works include evaluation of the time factor of detecting spams at an early stage using the identified user-behaviors features.

6. Acknowledgement

This Research was funded by the TETFund Research Fund and the Africa Centre of Excellence, Obafemi Awolowo University (OAU) - ICT Driven Knowledge Park.

References

Al-Zoubi, A.M., Alqatawna, J., and Faris, H. (2017). Spam profile detection in social networks based on public features. *In proceedings of the 2017 8th International Conference on information and Communication Systems (ICICS),* 130-135.

Amitay, E., Carmel, D., Darlow, A., Lempel, R., and Soffer, A. (2003). The connectivity sonar: Detecting site functionality by structural patterns. *In Proceedings of the 14th ACM Conference on Hypertext and Hypermedia, ACM, New York,* 38–47.

Baeza-Yates, R., Boldi, P., and Castillo, C. (2006). Generalizing pagerank: damping functions for link-based ranking algorithms. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'06, DOI: 10.1145/1148170.1148225.

Becchetti, L. Castillo, C., Donato, D., Leonardi, S., and Baeza-Yates, R. (2006). Using rank propagation and probabilistic counting for link-based spam detection. In Proceedings of the Workshop on Web Mining and Web Usage Analysis, WebKDD'06, 2006.

Bencz´ur A., Csalog´any K., and Sarl´os, T. (2006). Link-based similarity search to fight Web spam. In Proceedings of the Second Workshop on Adversarial Information Retrieval on the Web, AIRWeb'06.

Castillo, C. and Davison, B. (2011). Adversarial Web search. *Foundations and Trends in Information Retrieval Journal*, 4(5), 377-486.

Castillo C., Donato D., Gionis, A., Murdock, V, and Silvestri, F. (2007). Know your neighbors: Web spam detection using the web topology. *In Proceedings of the 30th Annual International ACM SIGIR Conference (SIGIR),* 423–430.

Caverlee J. and Liu L. (2007) Countering web spam with credibility-based link analysis. *In Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing, PODC'07,* 157-166, Doi>10.1145/1281100.1281124

Davison, B. (2000). Recognizing nepotistic links on the Web. *In Proceedings of the AAAI-2000 Workshop on Artificial Intelligence for Web Search*. 23–28.

Erd´elyi, M., Garz´o, A., and Bencz´ur, A. A. (2011). Web spam classification: a few features worth more. *In Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality, WebQuality'11, Hyderabad, India, DOI: 10.1145/1964114.1964121*

Fetterly, D., Manasse, M., and Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam Webpages. *In Proceedings of the 7th International Workshop on the Web and Databases*. 1–6.

Gyongyi, Z. Garia-Molina, H. and Pedersen J. (2004). Combating web spam with trustrank. *In Proceeding of the Thirtieth international conference on Very large data bases - VLDB '04*, 30, 576-587.

Gy¨ongyi, Z. and Garcia-Molina H. (2005). Web spam taxonomy. *In B. D. Davison (Ed.), Proceedings of the First International Workshop on Adversarial Information Retrieval (AIRWeb),* 39–47.

Hochbaum, D., Spaen, Q. and Velednitsky, M. (2019). Detecting Aberrant Linking Behavior in Directed Networks. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, 72-82, DOI: 10.5220/0008069600720082. Idowu, P. A., Egejuru, N. C., Balogun, J. A., and Sarumi, O. A. (2019). Comparative Analysis of Prognostic Model for Risk Classification of Neonatal Jaundice using Machine Learning Algorithms. *Computer Reviews Journal*, 3, 122-146. Retrieved from https://purkh.com/index.php/tocomp/article/view/312

Jindal, N., and Liu, B. (2008). Opinion spam and analysis. *In Proceedings of the ACM 2008 international conference on web search and data mining*, 219-230.

Kolhe, M. and Bhukte, D. (2015). Data Mining for Web Spam Detection Analysis of Techniques. *International Journal of Science and Research (IJSR),* 5(10), 1395 – 1399.

Liu, Y., and Chen, R., Zhang, M., Ma, S. and Ru, L. (2008). Identifying Web Spam with User Behaviour Analysis. *In Proceedings of the 4th international workshop on Adversarial information retrieval on the web AIRWeb '08*, 9-16, doi>10.1145/1451983.1451986

Manne, G. and Wright, J. (2011). If Search Neutrality is the Answer, What's the Question? *SSRN eLibrary.* 10.2139/ssrn.1807951.

Mishne, G., Carmel, D. and Lempel, R. (2005). Blocking blog spam with language model disagreement. *In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, AIRWeb'05, Chiba, Japan.*

Najork M. (2009) Web Spam Detection. *In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, https://doi.org/10.1007/978-0-387-39940-9.*

Ndumiyana, D. & Magomelo, M. and Sakala, L. (2013). Spam Detection using a Neural Network Classifier. *Online Journal of Physical and Environmental Science Research*, 2, 28-37.

Niu, X., Liu, G. and Yang, Q. (2018). Trustworthy Website Detection Based on Social Hyperlink Network Analysis. *IEEE Transactions on Network Science and Engineering,* 1-12, DOI 10.1109/TNSE.2018.2866066.

Ntoulas, A. Na jork, M. Manasse, M. and Fetterly, D. (2006). Detecting spam web pages through content analysis. *In Proceedings of the 15th international conference on World Wide Web*, 83-92. doi>10.1145/1135777.1135794

Piskorski, J., Sydow, M., and Weiss, D. (2008). Exploring linguistic features for web spam detection: a preliminary study. *In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, AIRWeb'08, Beijing, China,* 25-28, Doi>10.1145/1451983.1451990.

 DOI: 10.7176/NCS

Rao, A.S., Avadhani, P.S. and Chaudhuri, N.B. (2016). A Content-Based E-Mail Filtering Approach using Multi-Layer Perceptron Neural Networks. *International Journal of Engineering Trends and Technology (IJETT),* 41(1), 44 – 55.

Román, P.E., Dell, R.F., Velásquez, J.D. and Loyola, P.S. (2014). Identifying user sessions from web server logs with integer programming, Intel. Data Anal. 18 (1), 43–61

Spirin N. and Han H. (2011). Survey on web spam detection: principles and algorithms. *ACM SIGKDD Explorations Newsletter archive*, 13(2), 50-64. doi>10.1145/2207243.2207252.

Svore, K. M. Wu, Q., Burges, C. J. C. and Raman, A. (2007). Improving web spam classification using ranktime features. *In Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web, AIRWeb'07, Banff, Alberta,* 9-16, doi>10.1145/1244408.1244411.

Sydow, M., Piskorski, J., Weiss, D. and Castillo, C. (2007). Application of machine learning in combating web spam, 2007.

Wei, C., Liu, Y., Zhang, M., Ma, S., Ru, L.' and Zhang, K. (2012). Fighting against Web Spam: A Novel Propagation Method based on Click-through Data. *SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval,* 395-404

Wu B. and Davison B. (2006). Detecting semantic loaking on the web. *In Proceedings of the 15th International Conference on World Wide Web, WWW'06,* 819-828.