Journal of Theoretical and Applied Information Technology
31st May 2022. Vol.100. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HYBRIDIZED SPAMBOT DETECTION SYSTEM USING SOURCE-CONTENT CLASSIFICATION

ISMAIL ADELABU ADEGBOLA[1], OLAIYA FOLORUNSHO[2,3], TAOFEEKAT TOSIN SALAU-IBRAHIM[4], RASHEED GBENGA JIMOH[5]

[1]Department of Computer Science, Oyo State College of Education, Oyo State, Nigeria
[2]Unit for Data Science and Computing, School of Computer Science and Information Systems, North-West University, Potchefstroom 2531, South Africa
[3]Department of Computer Science, Federal University Oye Ekiti, Nigeria
[4]Department of Computer Science, Al-Hkimah University, Ilorin, Nigeria
[5]Department of Computer Science, University of Ilorin, Ilorin, Nigeria

E-mail: [1]ismailadegbolaa@gmail.com, [2,3]olaiya.folorunsho@fuoye.edu.ng, [4]ttsalau@alhikmah.edu.ng, [5]jimoh_rasheed@yahoo.com

ABSTRACT

A web robot is an automated program that requests web resources independently from a web server. The growth of bad and good web robot traffic on web 2.0 over the years is on the high side compared to humans. A malicious web robot has been used to spam activities in web 2.0. It, however, poses serious challenges to website owners and can cause Distributed Denial of Service (DDoS) attacks. Previous web robot detection techniques emphasized pre-processing access-log files for identifying web robot session. A few approaches considered the semantic analysis of the content requested within a session as a source of web robot detection. Furthermore, very little effort has been made in combining the strength of behavioural features and semantic features in web robot detection. Therefore, this paper aimed at developing hybridized spambot detection system using source–content classification. This paper revealed that 7568 unique sessions were identified with 6993 humans, 558 spambots and 17 non-spambots; session text coherence (STC), session word relatedness (SWR) and session topic coherence (ST) were functionally expressed as $STC = sw/(n * m)$, $SWR = k/(n * m)$, and $ST = c/(n * m)$ respectively, where n represents the number of relevant topics, m number of top words in each topic, sw sum of weights, c count of unique topics and k count of unique word. The hybridized features performed effectively with an accuracy of 93.67% compared to the individual features.

Keywords: *Source, Content, Spambot, Detection, Web 2.0.*

1. INTRODUCTION

Web robot is an automated computer program (script) that requests web resources independently in web 2.0 from a web server [1]. The growth of bad and good web robot traffic on web 2.0 over the years is on the high side compared to that of a human user. This is evident in the [2] report, which shows 9.5% growth in bad web robot from previous year, 8.8% growth of good web robot from the previous year and a decrease of -5.8% in human traffic of the prior year. Spambot has been the major tool used to cause distributed denials of service (DDoS) attacks to information system security on the internet today. However, the presence of web robot cannot be eradicated totally due to their beneficial roles such as websites indexing, search engine optimization, automating routine activities and so on.

Just as normal email spam reduces the integrity and legitimacy of an email, an unsolicited comment posted to a dynamic site renders the page useless and exposes the website to being blacklisted [3]. So many websites are at the crossroad of losing customers or visitors since their integrity and legitimacy of their content call for concern [4]. Humans normally browse the web for a piece of information in a unique topic; whereas most of the web robots surfed uniformly with no priority to a page or content [5]. Semantic (in) coherent of the content requested by the user is necessary to determine a session that is induced by a web robot.

Previous web robot detection techniques emphasized a behavioural approach by pre-

processing access-log files to identify unique features for web robot sessions [6]. Recent detection assumed that humans typically look for specific information on a particular topic; on the other hand, most web robots go through the content of websites in a uniform fashion, without favouring specific pages or content [1].

This approach demand semantically measuring the coherence of the content requested. The semantic approach is necessary because web 2.0 is very rich in content, and a few of the approaches considered the semantic analysis of the content requested within a session as a source of web robot detection [5]. Furthermore, very little effort was made to combine the strength of behavioural features and semantic features for hybridised web robot detection in content-rich websites. Therefore, this paper formulated and evaluated three semantic features set and combined the existing behavioural features and semantic features formulated for web robot detection. The outcome of this research work will be used to build a robot detection mechanism that will reduce the interaction of spambot and consequently increase the integrity of the content of a dynamic website.

2. WEB ROBOT DETECTION PROBLEM

The problem of web robot detection can be perceived as being a classification problem [7-9]. Given a session S with n number of requests for web pages (overall number of computer documents that form a site text corpus). We label whether the session is induced by a web robot such as Spambot, indexers, auto filler, or an interactive user. A typical HTTP request from a server will have, but not limited to the following fields depending on the server arrangement: the IP address of where the requested is initiated, the date with time the request was made, the name of the resources requested, the HTTP method used and response code received by the client from the server, user agent string from the request packet that identifies the browser used, and the referrer field from which the user or client navigates to the resources.

Web robot detection has three levels of abstraction – the highest level is IP address, followed by the user, and the lowest level is the session. The session-level seems to be the point of focus for past researchers in web robot detection. This is stemming from the fact that IP level can have more than one user and the user level group can contain more than one session. The session-level includes information on user browsing for a unique visit. The past research posited that session-level can be modelled and analysed quickly when compared to IP and user-level [10].

A session refers to a group of HTTP requests received by a web server from the user in one visit. This definition implies a lot of pre-processing of the raw server access log to reflect user visits during a browsing session. Therefore, further processing can now be employed on the session logs by extracting various features and applying a technique to determine if it belongs to a human or web robot. The above definition applies to offline log analysis since the session has been terminated before detection is made. For real-time detection, the incoming stream of requests is analysed and determined whether a web robot or a human induces it before the session ends [11].

It is believed that real-time detections are more tedious since it has to work with few active session data, but the essence is to detect web robot as the session is going on. On the other hand, offline web log analysis offers a holistic approach to web robot detection because a large volume of data is involved in the process, reflecting the different patterns that can be modelled to detect unknown web robots.

This paper implements a hybrid of source-content classification approach to web robot detection. The source is the behavioural features as seen in the web access log and the content is the semantic features as seen in the incoherence of the content requested by the user. The approach proposed the use of weblog analysis in a hybrid format by examining the features of the raw log that can inherently differentiate web robot from a human session and the semantic incoherence of the requested resources by the user during a session. This approach is in line with the assertion proposed by [1] which is based on a notion that humans typically look for a specific topic while surfing the website, while web robot surf with no priority to a specific page. This assumption demands semantic measurement in terms of the incoherence of requested content within a session.

3. RELATED WORKS

The study of [12] advocated a navigational pattern approach for web robot detection, most especially a camouflaging and unknown web robot. The study identifies multiple IP addresses and user agents and introduced the idea of a threshold to determine the beginning and the end of a session that can solve the problem of a grouped IP/User-agent having more than one session. As shown in [13], classified commonly used method to detecting web robot into four. They include the simple methods, that is, matching the [agent] and [IP

address] with a known web robot, robots.txt file request checking, the use of traps method, that is, checking of embedded HTML code access which is similar to links that are invisible to human user and web navigation behaviour analysis method, which is the ability to find characteristics of the log that differentiate a web robot based on its objectives. They classified the web robot detection process using the web data pre-processing phase and the data mining phase.

The work of [14] investigated the spamming activities in web 2.0 and its findings indicated that spamming activities in web 2.0 are different from earlier static websites in construction and dissemination. Web 2.0 allows a non-technical person to host and publish information on their product or services on several web 2.0 applications such as sites used for social interactions, blogs, forums, wikis articles, etc. This allowed spammers to employ automated tools to create a fake profile, respond to a thread, manipulate wiki articles in an online forum in order to host and distribute spam content. [15] examined the spamming activities in the internet of things (IoT) domain. The author stressed that spamming activities could take place in the cyber world at any point in time. It may range from getting unsolicited or uninvited emails to the harmful or mischievous changes made to web pages by Web spambots. Internet of Things (IoT) is no exception for spamming activities as IoT provides an enabling environment for a physical object to be represented and used in the cyber world. The author examined how 2D barcodes are used to submerge the physical side of the Internet of Things, with links that lure the user to spam content and destroy legitimate content.

The study of [16] assumed that regardless of the function upon which a web robot is developed would display a focused and consistent behaviour as opposed to a human user. The author further revealed that web robots could use various search strategies, including depth-first and breadth-first algorithms. They proposed switching factors as a feature for detecting different types of web robots. [17] Recognizes the Distributed Denials of Service (DDoS) attack as being caused by a malicious web robot on the World Wide Web. The author analyzes the web user log using two neural network unsupervised algorithms. This is an attempt to examine the type and distribution of web users to a website using their browsing behaviour. The author further aims to obtain a useful insight into major differences and similarities useful web robot and a harmful web robot. In the study [18] the authors explore the use of a semi-supervised algorithm to detect a web robot. The author improves the Support Vector Machine as a classifier by retraining the classifier in an iterative manner.

The study of [1] examined the use of weblog entries and semantic analysis of the requested resources with a supervised machine learning algorithm to detect web robot in an academic publishing website. The author relies on the assumption that human users will be accessing the web for a specific article or domain whereas a web robot crawls the web incoherently. Building on this assumption, require topic modelling of the websites and semantic coherence using Latent Dirichlet Allocation (LDA). However, LDA generates less coherent topics and does not incorporate a supervised label set into its learning procedure. The work of [19] summited that LDA can over-generalize topics and offers no obvious way of incorporating a supervised label set into its learning procedure. This study implements Non-Negative matrix factorization (NMF). It was observed that top keywords of the topic NMF find are more related and meaningful to the corpus context than LDA.

4. SPAMBOT DETECTION FRAMEWORKS

The developed system commenced with the analysis of a web user log collected from a forum for three months. The study extracted two features from the log entries to the server: the source and content features. The source features are the behavioural features of log entries as recorded by the server and content features (semantic features) as reflected from the requested webpage. The two features were pre-processed to reflect user browsing sessions and further divided into 60% for training and 40% for testing.

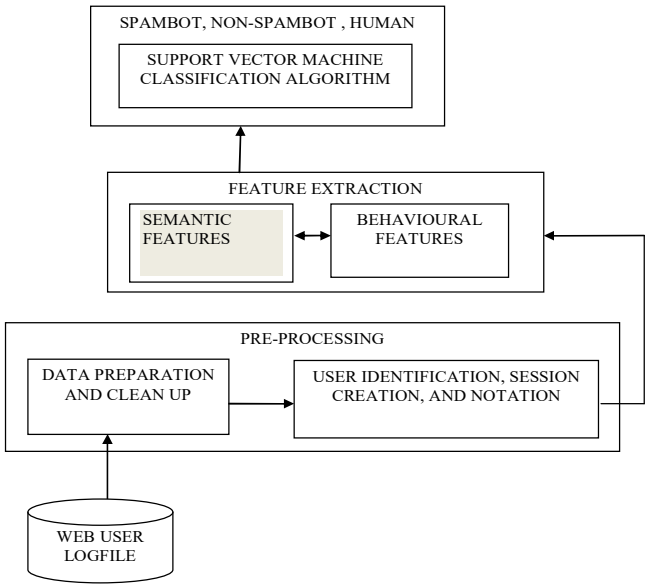

Figure 4.1: Spambot Detection Framework

4.1 Data Set Description

The study used the web user access log from a public online community forum. The log files contain the requests from a server for a period of three months. The data is further pre-processed to remove the login page, home, and other entries that cannot be used for semantic analysis due to its insufficient text length.

4.2 Data Labelling

The problem, being a multiclass problem, the study identified and labelled session as follows: Human, Spambot, and Non- Spambot. Human: This is an interactive user that browses the internet to access various resources and services provided by a web server.

Spambot: This is all types of web robot with the sole intention to distribute spam content and cause various forms of attacks such as a denial of service to a web server. Non Spambot: This refers to a specific web robot used to improve interactions with a web server. It is not primarily developed to cause any form of attack on a web server. Examples include Link checker and Auto filler.

4.3 SEMANTIC FEATURES EXTRACTION

This requires topic modelling and measurement of coherence of the content requested. The following features have been proposed to measure the coherence of the content requested within a session.

Given a session S with n number of requests for web pages (overall number of computer documents that form a site text corpus).

Let T be the number of relevant topics and W be the number of words at the top of each topic

i. **Session Text Coherence (STC):** This measure the degree of coherence among the documents/pages in a session. It is a numerical value calculated as follows:

STC = (sum of weights of words in the documents in session across all relevant topics) / (T x W); for a human, it is expected to have a higher value and a lower value for a bot.

Algorithm: Calculating Session Text Coherence (STC)

1. Let n be number of relevant topics
2. Let m be chosen the number of top words in each topic
3. Let T be set of Topics in the session and t_i is topic at ith position in set T
4. Let w_{ij} be the weight of the word at j_{th} position in topic t_i
5. Let sw be sum of weights =0
6. **for each** t in T **do**
7. sw =sw+w_{ij}
8. **end for**
9. **STC** = sw/(n x m)

Figure 4.2 Shows The Algorithm For Calculating Session Text Coherence.

ii. **Session Word Relatedness (SWR):** This is a numerical value calculated as follows: SWR= (unique count of words of session found in relevant topics) / (T x W).

Algorithm: Calculating Session WordRelatedness (SWR)

1. Let n be number of relevant topics
2. Let m be chosen the number of top words in each topic
3. Let T be set of Topics in the session and t_i is topic at ith position in set T
4. Let w_{ij} be the weight of the word at j_{th} position in topic t_i
5. Let c be count of unique words =0
6. Let U be set of unique words = {Ø }
7. **for each** t in T **do**
8. **if**w_{ij}not in U **then**
9. c = c+ 1
10. U_c= w_{ij}
11. **end if**
12. **end for**
13. **SWR** = c/(n x m)

Figure 4.3: Shows The Algorithm For Calculating Session Wordrelatedness (SWR)

iii. **Session Topic Coherence (ST):** This numerical value is calculated as follows: ST = (unique count of topics that exist in a session) /T without counting twice.

Algorithm: Calculating Topic Coherence (ST)

1. Let n be number of relevant topics
2. Let m be chosen the number of top words in each topic
3. Let T be set of Topics in the session and t_i is topic at ith position in set T
4. Let w_{ij} be the weight of the word at j_{th} position in topic t_i
5. Let c be count of unique topics =0
6. Let U be set of unique topics = {Ø }
7. **for each** t in T **do**
8. **if***t*not in U **then**
9. c = c+ 1
10. U_c= t
11. **end if**
12. **end for**
13. **STC** = c/(n x m)

Figure 4.4: Shows Algorithm For Calculating Topic Coherence (ST).

Algorithm : Semantic Feature Extraction

1. Let S be set of labelled session documents
2. Let V be set of semantic feature vectors = {Ø } such that vi is a feature vector for session I and vij is semantic feature value for attribute j
3. Let i be counter =0;
4. **for each** s in S **do**
5. i = i +1
6. vi[STC]= computeSTC(s.documents)
7. vi[SWR] =computeSWR(s.documents)
8. vi[ST] =computeST(s.documents)
9. **end for**

Figure 4.5 Shows The Algorithm For The Extraction Process Of The Semantic Features.

NMF ALGORITHM

Algorithm V: Building Topic Model for Semantic Feature Extraction

1. Let V be document-word matrix – input that contains which words appear in which documents of corpus.
2. Let W (Basis vectors)—the topics (clusters) discovered from the documents.
3. Let H (Coefficient matrix)—the membership weights for the topics in each document.
4. W and H be Initialized with non-negative factors
5. We calculate W and H by optimizing over an objective function (like the EM algorithm), updating both W and H iteratively until convergence.
6. **Do Until** W and H are stable: update the values in W and H by computing the following, with n as an index of the iteration.
   i. $$H_{[i,j]}^{n+1} \leftarrow H_{[i,j]}^{n} \frac{((W^n)^T V)_{[i,j]}}{((W^n)^T W^n H^n)_{[i,j]}}$$
   ii. $$W_{[i,j]}^{n+1} \leftarrow W_{[i,j]}^{n} \frac{(V(H^{n+1})^T)_{[i,j]}}{((H^{n+1})^T W^n H^{n+1})_{[i,j]}}$$
7. **end do**
8. nmf_model= W.H

Figure 4.6 Shows The Algorithm For Building Topic Model For Semantic Feature Extraction.

5. RESULTS AND DISCUSSION

5.1 Summary of Dataset

The study performed a replication on five randomly split datasets for behavioural features, semantic features and combined features. The predictive accuracy, Recall, Precision, and F-scores were measured for each dataset. Summary of the session created following the labelling technique that was explored is shown in Table 6:

7568 sessions were created comprising of 6993 Human session which is equivalent to 92%, 17 Non Spambot sessions and 558 Spambot sessions which is equivalent to 8% web robot. The dataset was divided into 60% for training and 40% for testing and this was further randomly split into five runs.

Table 5.1: Dataset Summary

Dataset	Human	Non-Spambot	Spambot	Total
Total Data	6993	17	558	7568
Training (60%)	4195	10	334	4539
Test (40%)	2798	7	224	3029

5.2 Experimental Result of Behavioural Features

Table 5.2 shows the experimental result on five random datasets for behavioural features used in spambot detection based on accuracy, error rate, recall, specificity, precision, and F-score.

Table 5.2: Experimental Results For Behavioural Features

Test Runs	Accuracy (%)	Error Rate (%)	Recall	Specificity	Precision	F1-Score
Test Run 1	92.12	7.88	0.987	0.466	0.922	0.953
Test Run 2	91.01	8.99	0.985	0.413	0.905	0.943
Test Run 3	91.81	8.19	0.986	0.435	0.921	0.952
Test Run 4	90.05	9.95	0.983	0.364	0.882	0.930
Test Run 5	90.17	9.83	0.984	0.408	0.897	0.938

From table 5.2 the average accuracy of 91.03% was achieved, ranging from 90.05 in Test Run 4 to 92.12 in Test Run 1. The highest accuracy was achieved by TestRun1 with error rate 7.88, recall of 0.987, and specificity of 0.466, the precision of 0.922 and an F1 score of 0.953. This implies the result is better and there is a high degree of certainty in the decision of the classifier.

5.3 Experimental Result of Semantic Features

Table 5.3: shows the experimental result on five random datasets for semantic features used in spambot detection based on accuracy, error rate, recall, specificity, precision, and F-score.

Table 5.3: Experimental Result For Semantic Features

Test Runs	Accuracy (%)	Error Rate (%)	Recall	Specificity	Precision	F1- Score
Test Run 1	88.45	11.55	0.979	0.362	0.892	0.933
Test Run 2	88.10	11.9	0.977	0.332	0.879	0.925
Test Run 3	88.20	11.2	0.981	0.345	0.893	0.935
Test Run 4	87.65	12.3	0.977	0.298	0.853	0.912
Test Run 5	87.90	12.1	0.975	0.314	0.859	0.913

From Table 5.3 the average accuracy of 88.06% was achieved, ranging from 87.65 in TestRun 4 to 88.45 in TestRun 1. The highest accuracy was achieved by TestRun1 with error rate 11.45, recall of 0.979, and specificity of 0.362, the precision of 0.892 and F1 Score of 0.933. This implies semantic features contribute to web robot detection and the decision is good.

5.4 Experimental Result of Combined Features

Table 5.4 shows the experimental result on five random datasets for the combined features used in spambot detection based on accuracy, error rate, recall, specificity, precision, and F-score.

Table 5.4: Experimental Result For Combined Features

Test Runs	Accuracy (%)	Error Rate (%)	Recall	Specificity	Precision	F1 Score
Test Run 1	93.67	6.33	0.996	0.66	0.961	0.978
Test Run 2	93.13	6.87	0.996	0.59	0.948	0.971
Test Run 3	93.34	6.66	0.995	0.64	0.961	0.977
Test Run 4	92.64	7.36	0.995	0.54	0.933	0.963
Test Run 5	92.7	7.3	0.995	0.58	0.941	0.967

From table 5.4 above, the average accuracy of 93.09% was achieved, ranging from 92.64 in Test Run 2 to 93.67 in TestRun 1. The highest accuracy was achieved by TestRun1 with error rate 6.33, recall of 0.996, and specificity of 0.66 and a precision of 0.961. The slight increment in the accuracy and corresponding recall value indicated that when the behavioural feature and semantic features are combined in web robot detection, there is better classification accuracy.

Figure 5.1 Shows A Comparison Of Accuracy Of Basic Features, Semantic Features And Combined Features.

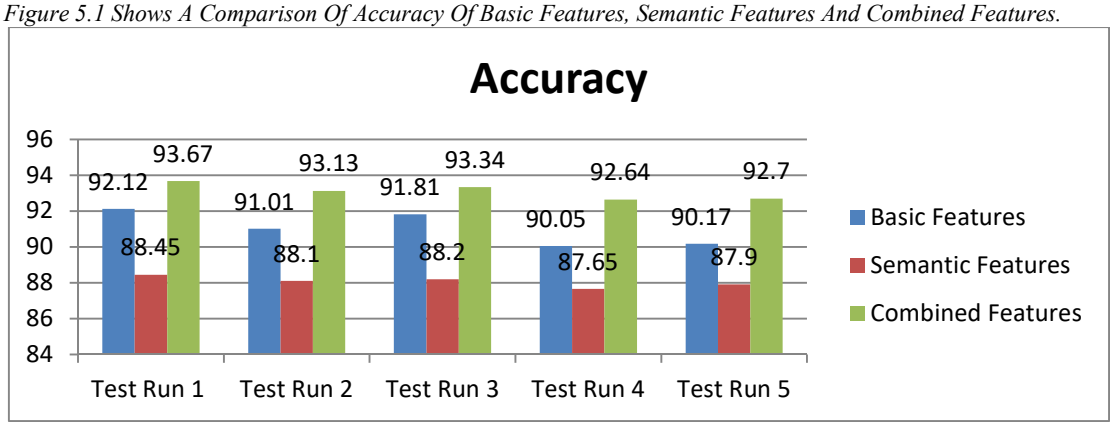

Figure 5.1: Accuracy Comparison

In figure 5.1, it was observed that when the basic features and semantic features are combined, the classification algorithm increases significantly. This has shown that semantic analysis of the user session is another good way of detecting spambot.

Figure 5.2: Comparison Of Recall Score Of Basic Features, Semantic Features And Combined Features

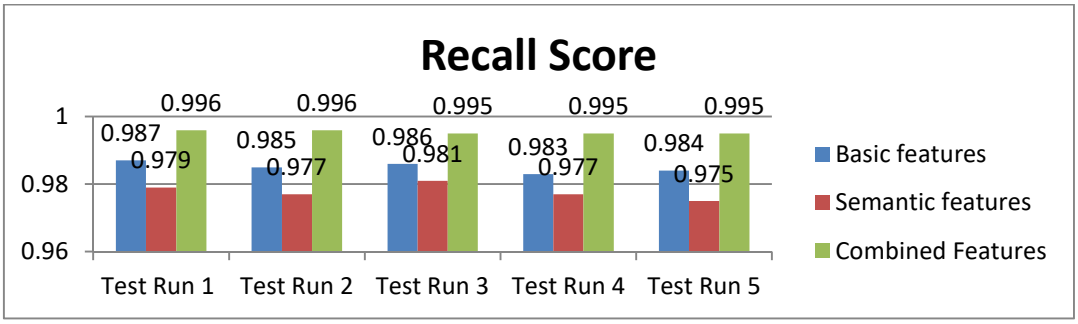

Figure 5.2: Recall comparison

In figure 5.2, when the basic features and semantic features are combined, the Recall or sensitivity of the classification algorithm increased significantly. This has shown that semantic features implemented are useful in analysing user session to detect spambot.

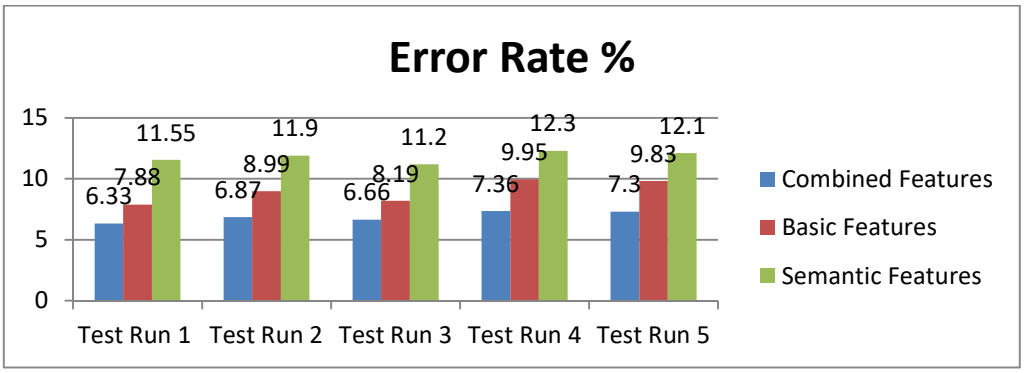

Figure 5.3: Error Rate Comparison

In figure 5.3, it was observed that the combined features have the least error rate, follow by the basic feature and the semantic features. Figure 5.4: Comparison of Specificity of Basic Features, Semantic features and Combined Features

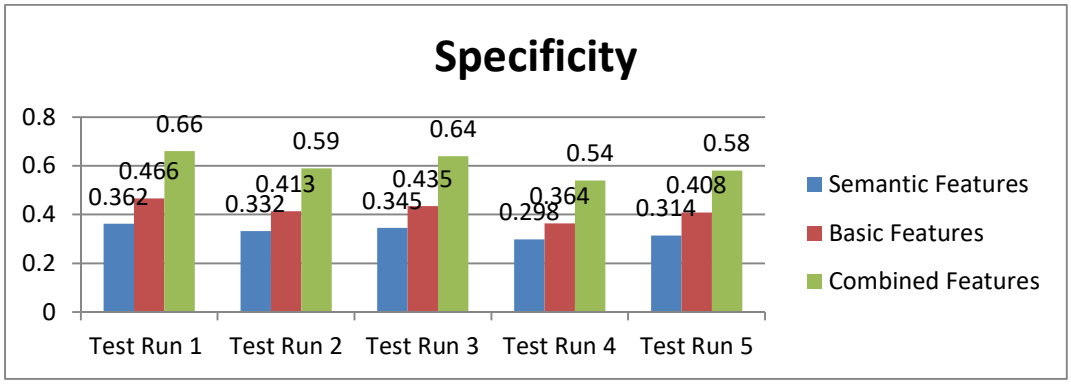

Figure 5.4: Comparison Of Specificity

In figure 5.4, it was observed that the semantic features have the least specificity followed by the basic feature and the combined features.

Figure 5.5: Comparison of Precision of Basic Features, Semantic features and Combined Features

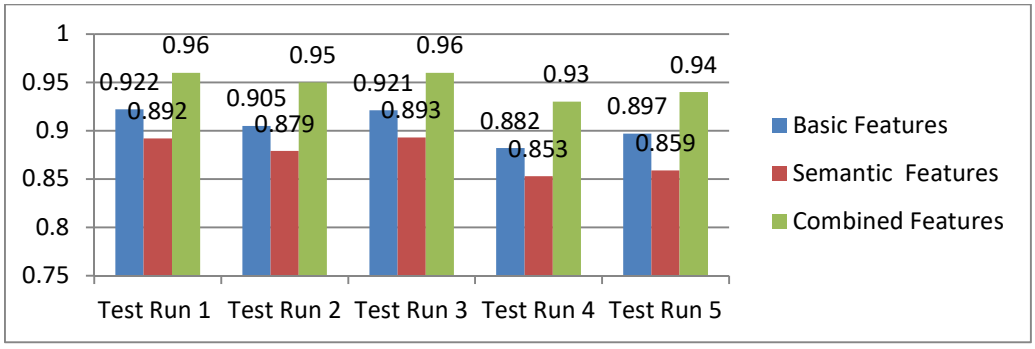

Figure 5.5: Comparison Of Precision

In figure 5.5, it was observed that the combined features have the highest precision followed by Basic features and semantic features.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. DISCUSSION OF FINDINGS

The study of [20] believed that it is difficult to see a technique as ideal because it depends on the tools used and operating procedure which may be difficult to implement faithfully and on the dataset used for experimentation. The best that can be done is to compare the strengths and weaknesses of their detection philosophy. The result of Support Vector Machine classifier revealed that the spambot was detected with an average accuracy of 91.03% with behavioural features alone. An average accuracy of 88.06% was obtained when semantic features was used alone. However, an average accuracy of 93.67% was recorded when the features were combined. This result showed that there was an increase in the accuracy, recall and precision when the features were combined. The implication is that the combined system performed better than the individual system and a high level of certainty was observed in the classifier's decision when the features were combined. This result is better than the result obtained by [1] whose accuracy was 90.07% when behavioural features were used, 84.84% when semantics features were used and 91.33% when the features were combined.

Table 6.1 compares the performance of the developed system with the existing system that uses the combination of source (Behavioural) and content (Semantic) approach to spambot detection in terms of accuracy, recall, and precision f1-score. It was discovered that the developed hybridized source-content system had a better performance when compared with the existing system.

Table 6.1: Combined Performance Evaluation

Author	**Approach**	**Accuracy**	**Recall**	**Precision**	**F-score**
[1]	Behavioural and Semantic approach with (LDA) as Topic Model	91.33	*	*	0.918
Developed Hybridized Spambot System	Behavioural and Semantic approach with NMF as Topic Model and New Features, STC, SWR, STC	93.67	0.996	0.961	0.978

7. CONCLUSION

The approach presented in this paper is a hybrid classification of web traffic using source-content classification. This approach relies on the behavioural features and semantic analysis of the content requested within a session as a means of web robot detection. The research seeks to confirm the assumption that a web robot, while navigating through the web, does so uniformly without prioritizing which page or content to access. Therefore, the paper introduces three features to measure the (in) coherence of the content requested within a session in terms of semantic. It also uses the behavioural feature of the session already existed in the literature and their relationship to detect web robots. There was an evaluation of the approach using the behavioural features from the source log alone, then the semantic features from the content requested within a session and then combining both features for web robot detection.

The approach presented offers two main advantages over the past approaches:

(1) Introduction of three coherent features that measure every aspect of the content of the web page. These include session text coherence, session word relatedness and session topic coherence.

(2) Implementation of a hybrid approach of behavioural approach and semantic analysis of the content requested within a session as a viable way of detecting web robot.

(3) Provide a means for classification of beneficial web robot called Non-spambot.

8. FUTURE WORK

In the future, researchers intend to compare the result of different classifiers with a non-probabilistic approach to topic modelling. Further analysis will be done in terms of the use of different SVM kernel function. Also researcher intends to work on behavioural and semantic feature selection.

REFERENCES:

[1] Lagopoulos, A., Tsoumakas, G., & Papadopoulos, G. (2017). Web robot detection in academic publishing. *arXiv preprint arXiv:1711.05098*.

[2] Globaldot Industry Report: Bad Bot Landscape 2018.Retrieved August 14, 2018 from https://www.globaldots.com.

[3] Hayati, P., Potdar, V., Chai, K., & Talevski, A. (2010). Web spambot detection based on web navigation behaviour. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications,* pp. 797-803.

[4] Adegbola, I., & Jimoh, R.. (2014). Spambot detection: A review of techniques and trends. *International Journal of Applied Information Systems, 6(9), 7-10.*

[5] Lagopoulos, A., Tsoumakas, G., & Papadopoulos, G. (2018, November). Web robot detection: A semantic approach. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI),* pp. 968-974.

[6] Stevanovic, D., Vlajic, N., & An, A. (2013). Detection of malicious and non-malicious website visitors using unsupervised neural network learning. *Applied Soft Computing*, *13*(1), *698-708.*

[7] Doran, D., & Gokhale, S. S. (2011). Web robot detection techniques: Overview and limitations. *Data Mining and Knowledge Discovery*, *22*(12), 183-210.

[8] Stassopoulou, A., & Dikaiakos, M. D. (2009). Web robot detection: A probabilistic reasoning approach. *Computer Networks*, *53*(3), *265-278*.

[9] Wohlkinger, W., & Vincze, M. (2010). 3D object classification for mobile robots in home-environments using web-data. *In 19th International Workshop on Robotics in Alpe-Adria-Danube Region,* pp. 247-252.

[10] Hayati, P., Chai, K., Potdar, V., & Talevski, A. (2010). Behaviour-Based web spambot detection by utilising action time and action frequency. In *International Conference on Computational Science and Its Applications, pp. 351-360.*

[11] Derek Doran and Swapna S Gokhale. 2016. An integrated method for real time and online web robot detection. Expert Systems 33, 6 (2016), 592–606.

[12] Tan, P. N., & Kumar, V. (2004). Discovery of web robot sessions based on their navigational patterns. *In Intelligent Technologies for Information Analysis 6(1) pp. 193-222*.

[13] Bomhardt, C., Gaul, W., & Schmidt-Thieme, L. (2005). Web robot detection-preprocessing web logfiles for robot detection. In *New developments in classification and data analysis* (pp. 113-124). Springer, Berlin, Heidelberg.

[14] Hayati, P., Chai, K., Potdar, V., & Talevski, A. (2009). HoneySpam 2.0: Profiling web spambot behaviour. In *International Conference on Principles and Practice of Multi-Agent Systems, pp. 335-344.*

[15] Razzak, F. (2012). Spamming the internet of things: A possibility and its probable Solution. *Procedia computer science*, *10, 658-665.*

[16] Kwon, S., Oh, M., Kim, D., Lee, J., Kim, Y. G., & Cha, S. (2012). Web robot detection based on monotonous behavior. *Proceedings of the information science and industrial applications*, *4*, *43-48*.

[17] Stevanovic, D., Vlajic, N., & An, A. (2013). Detection of malicious and non-malicious website visitors using unsupervised neural network learning. *Applied Soft Computing*, *13*(1), 698-708.

[18] Wang, D., Xi, L., Zhang, H., Liu, H., Zhang, H., & Song, T. (2015, August). Web robot detection with semi-supervised learning method. In *3rd International Conference on Material, Mechanical and Manufacturing Engineering (IC3ME 2015)*. Atlantis Press.

[19] Alsaedi, N. (2017). *Event identification in social media using classification-clustering framework* (Doctoral dissertation, Cardiff University).

[20] Doran, D., & Gokhale, S. S. (2016). An integrated method for real time and offline web robot detection. *Expert Systems*, *33*(6), 592-606.