Proceedings of the 37th iSTEAMS Cross-Border Conference – Accra Ghana 2023

Faculty of Computational Sciences & Informatics - Academic City University College, Accra, Ghana
SMART Scientific Projects & Research Consortium (SMART SPaRC)
Sekinah-Hope Foundation for Female STEM Education
ICT University Foundations USA
IEEE Computer Society, Nigeria Section
Intelligent Antitheft Management Networks
Creative Research Publishers – Society for Multidisciplinary & Advanced Research Techniques (SMART) Africa

37th International Science Technology Education Arts Management & Social Sciences (iSTEAMS) Cross-Border Conference - Accra Ghana 2023

# On The Use of Martin Porters Stemming Algorithm for Text and Image Spam Filtering For OCR and Word Clusters

**Oluwatunde, S.J., Gbolagade, K.A. & Asaju-Gbolagade, A.W.**
Department of Computer Science, Kwara State University, Malete, Kwara State, Nigeria
Department of Computer Science, University of Ilorin, Ilorin, Kwara State, Nigeria
**E-mails:** oluwatundesola@yahoo.com; kazeem.gbolagade@kwasu.edu.ng; gbolagade.aaw@unilorin.edu.ng
**Phone Nos.** +447909365912; +2348109668798; +2347067009668

## ABSTRACT

Techniques to detect text based spam mails have evolve over time and has become effective and scalable. To beat these systems, spammers have turned to image-based spams, taking advantage of the difficulties experienced by computers in extracting text from images. This difficulty arises because of the fact that computers can only solve well-defined problems, and images by their very nature are not well defined. Spammers further exploit this inherent weakness by obfuscating images even further. Several solutions involving the use of object character recognition have been proposed but with limited success. Against this background, we propose a research that will mathematically modelled a system that attempts to mitigate spam (text and image) by using spam words characteristic. For image spam, our model is expected to compensate for errors, which occurs as a result of the character recognition process by assigning the same probability to both the image text and its skewered text equivalence. In this paper, we present the research direction with a view to harvesting important inputs to the research progression.

**Keywords:** Martin Porters, Stemming Algorithm, Text, Image Spam, Filters, OCR, Word Clusters

**Proceedings Citation Format**
Oluwatunde, S.J., Gbolagade, K.A. & Asaju-Gbolagade, A.W. (2023): On The Use of Martin Porters Stemming Algorithm for Text and Image Spam Filtering For OCR and Word Cluster. Proceedings of the 37th iSTEAMS Multidisciplinary Cross-Border Conference. 30th October – 1st November, 2023. Academic City University College, Accra, Ghana. Pp 177-186.
dx.doi.org/10.22624/AIMS/ACCRACROSSBORDER2023V2P37

## 1. BACKGROUND OF THE STUDY

The Internet has revolutionized the modern world and the numerous Internet based applications that get introduced these days add to the high levels of comfort and connectivity in every aspects of human life. It is estimated that, over 2.095 billion people worldwide use Internet for various purposes (Jacob, 2023; Internetworld, 2024) – ranging from accessing information for educational needs to financial transactions, procurement of goods and services. As the modern world is gradually becoming "paperless' with huge amount of information stored and exchanged over the Internet, it is imperative to have robust security measurements to safeguard the privacy and security of the underlying data (Natarajan & Lopamudra, 2022; Naeem Ahmed; 2022).

Cryptography techniques have been widely used to encrypt the plaintext data, transfer the ciphertext over the Internet and decrypt the ciphertext to extract the plaintext at the receiver side. However, with the ciphertext not really making much sense when interpreted as it is, a hacker or an intruder can easily perceive that the information being sent on the channel has been encrypted and is not the plaintext. This can naturally raise the curiosity level of a malicious hacker or intruder to conduct cryptanalysis attacks on the ciphertext (i.e., analyze the ciphertext through the encryption algorithms and decrypt the ciphertext completely or partially) (Guilherme, Mirko & Ludovico, 2024).

The needs arise for more prudent and secured way to send the confidential information, either in plaintext or ciphertext, by cleverly embedding it as part of a cover media such as an image, audio or video file in such a way that the hidden information cannot be easily perceived to exist for the unintended recipients of the cover media. This idea forms the basis for Steganography, which is the science of hiding information by embedding the hidden (secret) message within other, seemingly harmless images, audio, video files or any other media. Steganography enables information transfer in a covert manner such that it does not draw the attention of the unintended recipients (Fumera, Pillai, & Roli (2020).

### 1.1 Steganograhic Technologies

Steganographic technologies are a very important part of the present and future of Internet security and privacy on open systems such as the Internet. Steganographic research is primarily driven by the lack of strength in the cryptographic systems on their own and many governments (especially in the developed countries) have created laws that either limit the strength of cryptosystems or prohibit them completely. This has been done primarily for fear by law enforcement not to be able to gain intelligence by wiretaps, etc. (Westfeld & Pitzmann, 2021).

This unfortunately leaves the majority of the Internet community either with relatively weak and a lot of the times breakable encryption algorithms or none at all. This is where steganography comes in, steganography can be used to hide important data inside another file so that only the parties intended to get the message even knows a secret message exists. To improve the security of the information to be hidden, it is a good practice to use cryptography and steganography together because both technologies together can provide a very acceptable amount of privacy in any system (Bret, 2022)

As internet comes under the reach of majority of the people the email becomes the cheapest and effective way of advertisement. Spammers are using this medium by sending unwanted email message through junk email, earlier textbased spam emails have been used but now to by-pass the conventional email filtering technique they are using imagebased spam. Image spam is actually a technique of embedding text (commercial content) into image by means of penetrating the text spam filter. Most email readers spend a non-trivial amount of time regularly deleting junk email messages, even as an expanding volume of such email occupies sever storage space and consume network bandwidth (Sahami, Dumais, Heckerman & Horvitz (2023). To protect the inbox from image spam emails, the filter should be able to distinguish between spam and ham images. The use of computer vision and pattern recognition techniques has been investigated in recent years and several text-based spam image filtering methods have been developed. Consequently some researchers proposed techniques based on detecting the presence of embedded text, and on characterizing text areas with low level feature like their size or their color distribution (Battista, Giorgio, Ignazio & Roli (2023).

However, some realistic problems that are not dealt well by the prevalent models. A spam image may belongs to several categories of spam images and similarity measurement is not able to discriminate because of small difference from each of the class. An intelligent spammer can send every time new image to defeat the spam filter as the result end user spam it after seeing it.

## 2. RELATED LITERATURE

### 2.1 The Evolution of Image-Based Spam

In 2003, the first image with the spam text inside was reported by GrahamCumming. Later, this technique was utilized successfully by spammers, by sending image spam as MIME attachments instead of sending as simple image tags. The previous content filtering techniques based on text analysis of subject and body fields of email were ineffective to handle this new spam attack type. The first attempts made by researchers to detect such spam were based on Optical Character Recognition (OCR) methods. These methods tried to extract the spam texts/words from image spam and compare with existing spam text keyword database (IBM, 2020; IBM 2021; Fabio, Battista, Giorgio, Ignazio & Riccardo Satta (2021); 2006; Longe & Chiemeke, 2012; Longe & Chiemeke, 2008; Longe, 2011)

Image-based spam or image spam is a kind of email spam where the textual spam message is embedded into images, that are then attached to spam emails. Since most of the email clients will display the image file directly to the user, the spam message is conveyed as soon as the email is opened (there is no need to further open the attached image file). The history of OCR can be traced all the way back to 1809, when reading devices for the blind or telegraph applications were developed. In 1914, Emanuel Goldberg developed a machine to convert printed characters into standard telegraph code. Concurrently, Edmund Fournier developed a handheld scanner called Optophone, which can produce tones corresponding to the specific letters/characters in the printed document (Herbert, 1982, Dalbe, 1914).

In the late 1920s, Emanuel Goldberg developed an optical code recognition system for searching microfilm archives. In 1950s, US Department of Defense created GISMO, a device that could read Morse Code as well as words on a printed page, one character at a time. In 1974, Ray Kurzweil developed omni-font OCR reading machine for the blind, which could recognize text printed in virtually any font (Li,; Li,; Xu,; Zhou,; Yan,; Xu & Zhang, (2018).

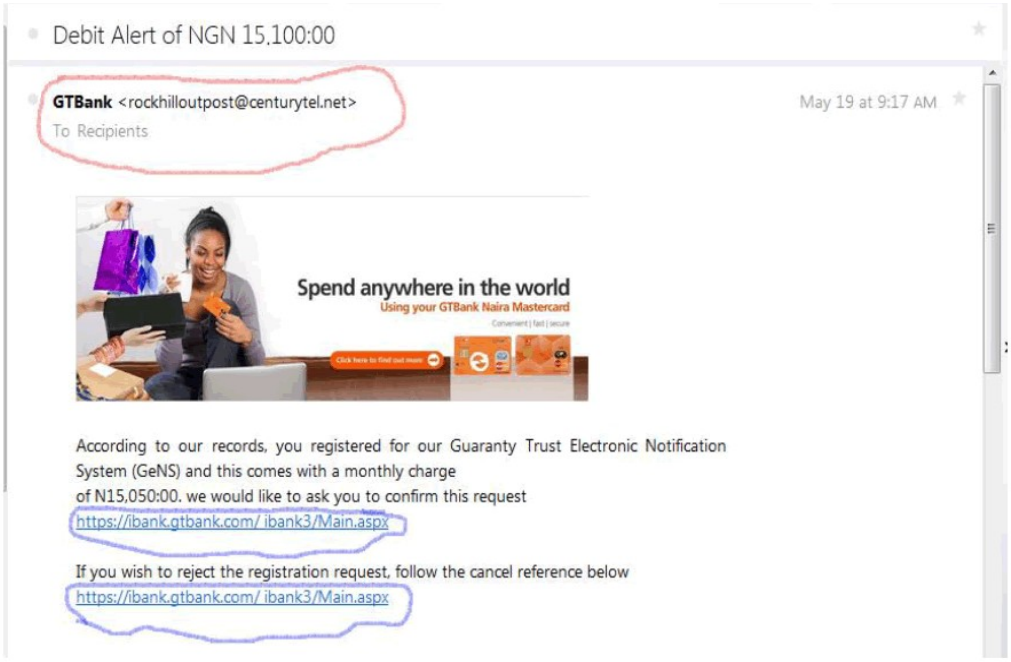


Fig 1.1: A Typical Image-based SPAM

**Source:** https://www.icommercecentral.com/open-access/internet-banking-identity-theft-and-solutions--the-nigerian-perspective.php?aid=86186

This device utilized the CCD flatbed scanner and the text-to-speech synthesiser. Later commercial version of the OCR computer programs were launched in the market for commercial purposes like uploading legal paper and news documents onto online databases. In the early 1990's, it was used by libraries for historic newspaper digitization projects. An open source GUI frontend PrintToBraille tool (Rose, 2009). was developed by A. G. Ramakrishnan and his team at Medical intelligence and language engineering lab, Indian Institute of Science, to convert scanned images of printed books to Braille books. Optical Character Recognition (OCR) is a pattern recognition technique which involves the process of converting the printed/typed or handwritten text (usually in the form of images) into machine-encoded text. Figure 1 shows the simple block diagram of OCR Reader. In data entry applications like passport documents, invoices, bank statements, computerized receipts, this technique is widely adopted. This method of digitizing printed texts helps to edit electronically the converted text and hence in turn, offers benefits like ease of searching, storing data in compact form, ease of displaying data on-line etc.

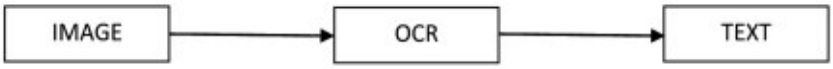


Fig 1: Block Diagram of an OCR Reader

The word steganography comes from the Greek Steganos, which means covered or secret and graphy means writing or drawing (Cole, 2023). Therefore, steganography means, accurately, covered writing. Steganography can be defined as the art and science of hiding information such that its presence cannot be detected (Bret, 2022). Just like the main goal of existing steganography system, this work improves the security of existing steganographic design and come up with more secured design that will make life difficult for steganalyst to break even if at all they detect it. The goal of image spam is clearly to circumvent the analysis of the email's textual content performed by most spam filters (e.g., SpamAssassin, RadicalSpam, Bogofilter, SpamBayes). Accordingly, for the same reason, together with the attached image, often spammers add some "bogus" text to the email, namely, a number of words that are most likely to appear in legitimate emails and not in spam. (Ebem, Onyeagba & Ugwuonah, 2018) The earlier image spam emails contained spam images in which the text was clean and easily readable, as shown in Fig. 1.1

## 2.2 Detection

Consequently, optical character recognition tools were used to extract the text embedded into spam images, which could be then processed together with the text in the email's body by the spam filter, or, more generally, by more sophisticated text categorization techniques.[3][6] Further, signatures (e.g., MD5 hashing) were also generated to easily detected and block already known spam images. Spammers in turn reacted by applying some obfuscation techniques to spam images, similarly to CAPTCHAs, both to prevent the embedded text to be read by OCR tools, and to mislead signature-based detection. Some examples are shown in Fig. 2.

This raised the issue of improving image spam detection using computer vision and pattern recognition techniques (**Battista, Giorgio, Ignazio & Fabio, 2018; Wu, Cheng, Zhu, & Wu, 2022**). In particular, several authors investigated the possibility of recognizing image spam with obfuscated images by using generic low-level image features (like number of colours, prevalent colour coverage, image aspect ratio, text area), image metadata. Notably, some authors also tried detecting the presence of text in attached images with artifacts denoting an adversarial attempt to obfuscate

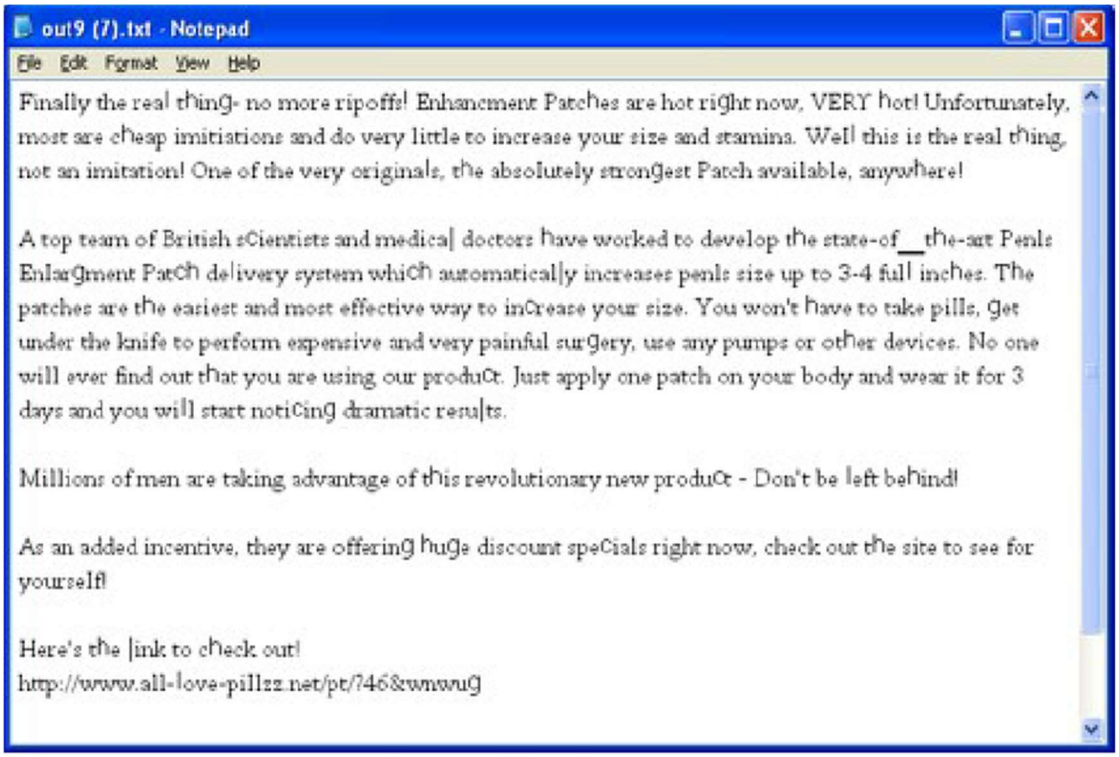


Fig 1.2: Obfuscated E-mail

Source: https://www.researchgate.net/figure/The-spam-message-after-obfuscation-In-this-obfuscated-email-some-characters-were_fig13_221462244

## 3. RESEARCH GAPS

The presence of unsolicited bulk emails, commonly known as spam, can seriously compromise normal user activities, forcing them to navigate through mailboxes to find the - relatively few - interesting emails. Even if a quite huge variety of spam filters has been developed until now, this problem is far to be resolved since spammers continuously modify their malicious techniques in order to bypass filters. In particular, in the last years spammers have begun vehiculating unsolicited commercial messages by means of images attached to emails whose textual part appears perfectly legitimate.

In this highly digitalized world, the Internet plays an important role for data transmission and sharing. However, due to its open nature, some confidential data might be stolen, copied, modified, or destroyed by an attacker. Therefore, security became an essential issue; the important of reducing a chance of the information being detected during the transmission is being an issue nowadays which led to the development of some security mechanism. Encryption is a well-known procedure for secured data transmission. Although encryption achieves certain security effects, they make the secret messages unreadable and unnatural (Guilherme, Mirko & Ludovico, 2024) but this protection can be broken with enough computational power.

An alternate approach to encrypting data would be to hide it by making this information look like something else and to hide the existence of the communication; this way only intended recipient would realize its true content. Existing Steganographic approach are still faced with the challenges of security and capacity. That is, they are easily detected by steganalysis and the amount of information they can effectively hide is low. The need to continuously develop systems that does not necessarily overwrite existing ones but rather improve security and ability to hide data for covet transmission is therefore warranted. This then is the thrust of this research

### 3.1 Research Direction

The wide use of image spam fetches the attention of researchers. Several attempts have been made to address filtering spam images by utilizing specific feature of image. For feature extraction there are various algorithms, such as principle component analysis (PCA), Independent component analysis (ICA), Partial Least squares (PLS) to transform graphical image into feature vector. In this paper, PCA has been used because of its suitability for data set in multiple dimensions.

Aradhye, Myers & Herson (2020) first addressed the grey mail problem and train two spam filters - the gray and b&w (ham/spam)filter on two disjoint subsets. Longe et al, 2021 proposed a maximum entropy using the portioned logistic regression (PLR) to learn content and user model separately. Despite these efforts, spammers are still able to beat these spam filters . In order to further optimize filtering efficiency and improve security, based on the foregoing, the aim of this research is to develop a system that Optimizes Text and Image Spam Filtering For OCR and Word Clusters Using Martin Porters Stemming Algorithm

The Specific Objectives are:

a. Carry out a Systematic Literature review of existing techniques with the view to evaluate existing mechanism for text and image-based spam filtering
b. Based on (a) above, to Identify the strength and weaknesses of the existing systems
c. Optimize and Design a more secured steganography system for text and image spam filtering for OCR and word clusters using Martin Porters stemming algorithm
d. Simulate the model the designed in (c) above;

## 4. SYSTEM MODEL

Steganography is derived from Greek and it means covered writing word and it is the art of communicating information between two parties in ways that remove the existence of the communication. For a communication to occur between a sender and a receiver, sender supply message M (which can either be in form of plaintext or file), and cover image C; combines them to generate a stego image Z which he sends to the receiver.

$$Z = f(M, C) \dots\dots\dots\dots eq.(1)$$

Function f in equation (1) is the system that combines message M and cover image C to generate the stego image Z which looks exactly the same as C. The method we are going to use in this research work is described as follows. Contrary to other methods used in hiding message in steganography, where message are hidden directly under the image, our method hides the encoded bits e, which is gotten from the function of the message M to be hidden and the image C to be used as the cover as shown in equation (2).

$$e = nC - nM \ldots\ldots\ldots\ldots eq.(2)$$

Note that our message M is the binary representation of the message. If the number n, of byte (8 bits) that our message contains is nM, we select the same n bytes (that is 8 bits per pixel because we are using grayscale as our cover image) from our cover image denoted with nC.

The resulting bit e from the equation (2) is embedded under the cover image using the least significant bit insertion method which is one bit per pixel until all the e bits has been inserted. The number of message bits n is converted to its binary equivalent and also inserted into the cover image from the rear pixel one bit per pixel. The essence of inserting n bit from the rear is to allow us note the number of bits to retrieve at the destination end whenever we want to collect out our encoded bits. This method improves the security of the hidden in the sense that we are not hiding the message directly under the cover file C, and every effort made by attacker to retrieve the message will result to a false message.

## 5. CONCLUDING REMARKS

This research work will allow users to encode message in a bitmapped image file in which the said image file will serve as the cover file which is a file that will house the intended message to be encoded and such that there will not be any discrepancies in the original cover file and the stego image (cover file + message) to ensure that the primary aim of steganography is not defeated. The research work is limited to the use of bitmapped image as its cover file and the size of message intended for encoding dictates the size of the cover file that can be used and vice-versa. It is expected that the outcome of this research work will significantly improve existing filters and further strengthen the transmission of covet text and image information. To implement the system, C# programming language and the Microsoft .NET Framework will be used.

## REFERENCES

1. Anjali A & Umesh L (2021), *A Secure Skin Tone Based Steganography Using Wavelet Transform*: International Journal of Computer Theory and Engineering, Vol. 10, No.1, February, 2021 1793-8201
2. Aradhye, H., Myers, G., Herson, J. A. (2020). Image analysis for efficient cat egorization of image-based spam e-mail. In: Proc. Int. Conf. on Document Analysis and Recognition, pp. 914–918.

3. Battista Biggio, Giorgio Fumera, Ignazio Pillai, Fabio Roli , "Image Spam Filtering Using Visual Information", 19th Int. Conf. on Image Analysis and Processing (ICIAP 2023), Modena, Italy, IEEE Computer Society, pp. 105–110, 10/09/2023.
4. Bret D (2022), *A detailed look at Steganographic Techniques and their use in an Open System Environment*: paper from the SANS Institute Reading Room site. SANS Institute, 2022.
5. Battista Biggio, Giorgio Fumera, Ignazio Pillai, Fabio Roli,*Biggio, Battista; Fumera, Giorgio; Pillai, Ignazio; Roli, Fabio (2018). "A survey and experimental evaluation of image spam filtering techniques, Pattern Recognition Letters". Pattern Recognition Letters. 32 (10):* Volume 32, Issue 10, 15 July 2018,
6. Cole E (2023), *Hiding in Plain Sight*: ISBN 416-236-4433 Wiley, John & Sons,
7. D. Ebem, Joseph Chinonye Onyeagba, G. Ugwuonah (2021): Internet Banking: Identity Theft and Solutions - The Nigerian Perspective. Law, Business, Computer Science. The Journal of Internet Banking and Commerce. https://www.semanticscholar.org/paper/Internet-Banking%3A-Identity-Theft-and-Solutions-The-Ebem-Onyeagba/54822b71a2044498d4c3b448e905497c5fcf3b8b
8. Fabio Roli, Battista Biggio, Giorgio Fumera, Ignazio Pillai, Riccardo Satta (2021) , "Image Spam Filtering by Detection of Adversarial Obfuscated Text", Workshop on Neural Information Processing Systems (NIPS), Whistler, British Columbia, Canada,
9. Fumera, G.; Pillai, I. & Roli, F. (2020): Spam filtering based on the analysis of text information embedded into images. Journal of Machine Learning Research (special issue on Machine Learning in Computer Security), 17:2699–2720, 2020.
10. Guilherme Ramos, Mirko Marras & Ludovico Boratto (2023): SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. https://sigir-2024.github.io/proceedings.html
11. Internet –www.wikipedia.com/internetworld, 2023
12. IBM X-Force® 2020, Mid-Year Trend and Risk Report (August 2020).
13. IBM X-Force® 2021, Mid-Year Trend and Risk Report (September 2021).
14. Https://Vijinimallawaarachchi.Com/2017/05/09/Porter-Stemming-Algorithm/
15. https://www.researchgate.net/publication/261394614_a_study_on_e-mail_image_spam_filtering_techniques
16. https://www.researchgate.net/publication/315388437_image_spam_filters_based_on_optical_character_recognition_ocr_techniques
17. https://www.virusbulletin.com/virusbulletin/2007/11/evading-spamassassin-obfuscated-text-images/Porter Stemming Algorithm – Basic Intro
18. Jacob Murel (2023): Stemming Text Using The Porter Stemming Algorithm In Python Use Watsonx, Nltk, And Spacy To Prepare Raw Text Data For Use In Ml Models And Nlp Tasks
19. https://developer.ibm.com/tutorials/awb-stemming-text-porter-stemmer-algorithm-python/*Li, Siyuan; Li, Ruiguang; Xu, Yuan; Zhou, Hao; Yan, Hanbing; Xu, Bin; Zhang, Honggang (2018-09-01). "WAF-Based Chinese Character Recognition for Spam Image Filtering". Chinese Journal of Electronics. 27 (5): 1050–1055. doi:10.1049/cje.2018.06.014. ISSN 1022-4653.*
20. Longe, O.B.(2011). On the use of Image-based Spam Mails as Carriers for Covert Data Transmission. Computer & Information Systems Journal. Vol. 15. Issue 1. http://cis.uws.ac.uk/research/journal/vol15.htm (DBLP Indexed)

21. Longe, O.B. & Chiemeke S.C. (2008). Probability Modeling For Improving Spam Filtering Parameters. Journal of Information Technology Impact. Vol 8, No.1. (SCOPUS, SCIMAGO Indexed)
22. Longe, O.B, Asani, E. & Ashante D. (2021): A Maximum Entropy Classification Scheme for Phishing detection using Parsimonious Features. TELKOMNIKA (Telecommunication Computing Electronics and Control), SCOPUS Indexed/ScimagoJR indexed journal, Q2 on Electrical and Electronics Engineering, SJR: 0.283, CiteScore: 1.09, SNIP: 0.730. Fabien P (2011), *The information hiding homepage Digital watermarking & steganography:* http://www.petitcolas.net/fabien/steganography/
23. Naeem Ahmed, Rashid Amin, Hamza Aldabbas, Deepika Koundal, Bader Alouffi & Tariq Shah (2022): Machine Learning Techniques For Spam Detection In Email And Iot Platforms: Analysis And Research Challenges. Https://Doi.Org/10.1155/2022/1862888. https://onlinelibrary.wiley.com/doi/10.1155/2022/1862888
24. Natarajan M and Lopamudra N (2022), *Steganalysis Algorithms for Detecting the Hidden Information in Image, Audio and Video Cover Media:* International Journal of Network Security & Its Application (IJNSA), Vol.12, No.1, 2022.
25. Sahami, M.; Dumais, S.; Heckerman, D. & Horvitz, E. (2023): A Bayesian approach to filtering junk e-mail. AAAI Technical Report WS-98-05, Madison, Wisconsin, 2023.
26. Westfeld A & Pitzmann A (2021), *Attacks on Steganography Systems*: Department of Computer Science, Dresden University of Technology
27. Wu, C.-T., Cheng, K.-T., Zhu, Q., Wu, Y.-L., 2022. Using visual features for anti-spam filtering. In: Proc. IEEE Int. Conf. on Image Processing, Vol. III.pp. 501–504.