FUPRE Journal
of
Scientific and Industrial Research

ISSN: 2579-1184(Print) ISSN: 2578-1129 (Online)

https://fupre.edu.ng/journal

Adaptive Hybrid Genetic Algorithm Trained Bayesian Network Framework for Spam Filtering Using Text Normalization

SUNDAY, O.[1], DAVID, N.[2], ANTHONY, I.[3], PATRICK, E.[4]

[1,2,3,4]Department of Computer Science, College of Communication & Telecommunications, Novena University, Ogume, Delta State, Nigeria

ARTICLE INFO

Received: 16/12/2022

Accepted: 27/03/2023

Keywords

Memetic algorithm, spam filters, semantic approach, text normalization, heuristic-processing, strings

ABSTRACT

The popularity of the short messaging services (SMS) has created a propitious environment for spamming to thrive. Spams are unsolicited advertising, adult-themed or inappropriate content, premium fraud, smishing and malware. They are a constant reminder of the need for an effective spam filter. However, SMS limitations of 160-charcaters and 140-bytes size as well as its being rippled with slangs, emoticons and abbreviations further inhibits effective training of models to aid accurate classification. The study proposes Genetic Algorithm Trained Bayesian Network solution that seeks to normalize noisy feats, expand text via use of lexicographic and semantic dictionaries that uses word sense disambiguation technique to train the underlying learning heuristics. And in turn, effectively help to classify SMS in spam and legitimate classes. Hybrid model comprises of text preprocessing, feature selection as well as training and classification section. Study uses a hybrid Genetic Algorithm trained Bayesian model for which the GA is used for feature selection; while, the Bayesian algorithm is used as classifier.

1. INTRODUCTION

The advent of short messaging services by Neil Papworth since 1992, has seen great penetration and a tremendous growth rate of the service. Advent of mobile phones with enhance features has contributed to the large scale adoption of SMS by users. The portability, mobility, ubiquity of services and its low cost continues to promote text messages to become the most used means of electronic communication in the world today (Ezpeleta et al., 2016; Ojugo & Eboka, 2020b). Short Message Service (SMS) is text service component of phones or mobile communication systems with standardized protocols that allow the exchange of short text messages between fixed line or mobile phone devices. An estimated 23-billion SMS is sent daily worldwide in 2014; While, a total of 8.3 trillion SMS was sent worldwide in the same year with net market revenue of over $128Billion in 2011. In 2016, the revenue was forecasted to be over $153Billion; And, evidence has shown that 3.39billion SMS was sent and received in Nigeria alone in the year 2013 (Ezpeleta et al., 2020; Ojugo & Eboka, 2018). The increased popularity and the consequent proliferation of SMS platforms have also seen a corresponding rise in unsolicited SMS called spam. The ITU 2005 campaign witnessed a rise in unsolicited commercial adverts as sent to mobile phones via SMS. Recent drift from email to SMS

Corresponding author: *kemmyadesanya@gmail.com*

DIO

spams is attributed to the availability of effective email filters, user awareness and industry collaboration (Ojugo & Oyemade, 2021).

Spams are unsolicited electronic messages that include emails, SMS, VoIP etc. They are unsolicited, messages from a sender – sent indiscriminately with no prior relations to a user mostly for commercial reasons (Yao et al., 2022; Yu et al., 2019). SMS spam range from adult-themed and inappropriate contents, unsolicited adverts, smishing and mobile-wares etc. SMS spams have since become enormous challenge – causing great loss of revenue to Internet Service Providers, Mobile Network Operators and users in general. The overall growth of spams grew by 300% from just 2011 to 2012 from millions of SMS received worldwide; And 33.3% attributed to spam-related messages. Also, in Nigeria alone, an estimated 334,857,685 SMS spam were received daily in 2015 (Sahmoud & Mikki, 2022). This implies that lots of mobile phone users are handicapped in the control of the number of spams they receive. Besides being distractive and annoying, users need a certain degree of privacy with their phones and free from Spam and viruses invasions (Akazue et al., 2022). Mobile network operators are geared towards reducing the number of spams over their network as such flooding makes the SMS channel more invasive and less secure (Hayati et al., 2010).

The tremendous rise in the usage of SMS is attributed to: (a) trust in SMS channel: SMS is a private communication between two parties only has created some level of trust and acceptance all over the world such that financial institution has adopted its use in payment authorization, (b) high open rate*:* Average time it takes to respond to SMS is faster than email and voice call – making it a preferred choice. Statistics have shown SMS has an average open rate of 99% and opens within 15-minutes; While, an email has an open rate of 20-25% and open with 24-hours, (c) low cost of transaction: Average cost per SMS is almost negligible, and free for some networks – affording mobile phone users the opportunity to send as many without recourse to cost. Marketers and many other institution has embrace bulk SMS a medium for advertising and interact with customers, and (d) ease and convenience of texting enables its use in nearly every environment without disrupting people around phone users; Unlike in voice call, SMS is in absolute silence without inconveniencing people around. Aided by the portable size of most mobile devices, communication can be done almost everywhere and any position (Jáñez-Martino et al., 2022; Lazarevic et al., 2005).

The SMS benefits both users and operators in diverse ways to include convenience, flexibility, seamless integration of message services and data access. Others benefits are: delivery notification, guaranteed delivery, reliable, low-cost for concise data, ability to screen messages and return calls, increased productivity, more sophisticated functionality and enhanced user benefits, delivery to multiple users at same time, ability to receive diverse information, e-mail generation, creation of user groups, integration with other data and Internet-based application, and increased revenue generation for the MNOs (Amalraj & Lourdusamy, 2022; Jáñez-Martino et al., 2020; Reads, 2014).

2. LITERATURE REVIEW

2.1 Sources and Consequents of Spam

The tremendous rise in the usage of SMS is attributed to its ease of use, ubiquity in nature, high open rates, low cost of transaction and inherent trust in the channel. The ease of use, portability, ubiquity, low open rate and low cost of SMS are major factors for its popularity and usage. This growth rate has equally attracted spamming to the channel. Spammers are well organized businesses seeking to make money through the use of email, mobile (SMS), instant message, social network and internet telephony channel without the consent of user (Ojugo & Eboka, 2019b, 2019a). Their merchandise are unsolicited advertising, inappropriate or adult-themed content, premium fraud, smishing and even distribution of malware generally called spam. SMS spams are thus, unsolicited and unwanted messages sent to mobile phone users. Spam trend is on the rise and its toll on subscribers and even MNO is getting intensive and proven to be of great concern to all (Jáñez-Martino et al., 2022; Redondo-Gutierrez et al., 2022).

Besides being quite distracting and annoying, spams have other serious consequences such as the competition for resources between millions of illegitimate and legitimate messages being transmitted. These messages consume network resources that could have otherwise been allocated to other legitimate services by MNO (Br et al., 2021). Spamming activities attracts extra cost for mobile operators to adequately maintain and service their mobile communication infrastructures for effective service delivery. Also flooding of MNO infrastructure with illegitimate massages can cause legitimate users to suffer denial of service. Huge amount of spam messages also concerns the cellular carriers as the messages traverse through the network, causing congestion and hence degrade network performance (Kamoru et al., 2017). Mobile communication industries are also faced with threat from virus, Trojan horse, worms and malware propagated by spam. Spam activities such as phishing, identity theft and other fraud related activities which were prominent in email messaging services has migrated to SMS platform (Karim et al., 2019; Karimi et al., 2013; Ojugo et al., 2021). Financial loss, damage to mobile user's reputation and that of the MNO are issues to be considered (Shelke & Sharma, 2020).

2.2 Types of Spam

SMS spam filters shares similar features and challenges with email spam filters. They are both saddled with the task of real-time filtering efficiency and the option to decide between client-side and or server-side filtering. The mobile space is faced with the challenge of overcoming misclassification cost and eliminate false-positives (genuine SMS incorrectly classified as spam by filter), and issue of concept drift in order to evade detection by system filters. Thus, most existing approaches of combating SMS spam are imported from successful email solutions (Nivedha & Raja, 2022; Ojugo & Eboka, 2019b, 2020a). But, not all solutions to email spam are applicable to SMS due to the fact that established email spam filters are unable to tackle SMS spam because the performance of email spam filters is seriously degraded when used to filter SMS spam. This is attributed to its limited 160-character of 140-bytes sized messages. Also, these messages are rife with slang, symbols, emoticons, and abbreviations that inhibit proper classification (Rathi & Pareek, 2013).

To overcome the shortfall of email filters in handling SMS spam successfully, a combination of filtering techniques to reduce noise in SMS and expands the message size – is the focus of this research. Spam filters can be divided into a number of broad categories based on the method used to filter Spam. They include (Redondo-Gutierrez et al., 2022; Sahmoud & Mikki, 2022) the following methods of filtering namely:

- ✓ List-based Filters – simply blocks spam (i.e. unwanted messages) using an already created list of senders. It then seeks to create a whitelist (i.e. a list of authentic senders), or a blacklist (i.e. a list of blocked sender records). Thus, when an incoming message is received – the spam filter simply checks if the IP, email address and/or mobile number is on the list. If the list is on the whitelist, the sender is accepted; otherwise, the message is not delivered to the user/subscriber. Vice versa in the case of a blacklist.
- ✓ Challenge and Response Filters – forces a message sender to prove themselves via series of tests. It blocks undesirable messages by forcing the sender to perform a task before their message is delivered. If the task is successful, the message (and future messages) are delivered to the

recipient; While, a failure to complete the challenge leads to message rejection.

- ✓ Content-based Filters – evaluates words or phrases found in messages to determine if the message is spam or not. It analyzes the message header, subject and body to discover any distinctive characteristic. It performs such feats via two-modes: word-based and heuristic filters.

  Word-based filters use a set of rules to detect genuine from spam SMS. Also known as rule-filters, they use rules about actual word(s) or phrase(s) in a message to classify messages into genuine and spam classes. Rule features include word type, frequency of occurrence, structure of text (e.g. font size, colour etc), presence of many periods between letters (e.g. F.R.E.E), existence of image, etc. Rules are filter-dependent and can vary from simple to very complex. A demerit of rule-based filters is that: (a) they are knowledge intensive, (b) time consuming process in reviewing spam messages to determine the rules, and (c) needs regular update of rules as spammers changes their tactics.

  Heuristic-filter examines data content via various algorithms and resources, and assigns points to words or phrases. Words commonly found in spams such as “FREE” or "SEX," receive higher scores. Terms commonly found in normal messages receive lower scores. The filter then adds up total scores. If the message receives a certain score or higher (determined by anti-spam application's administrator), the filter identifies it as spam and blocks it. Messages with score(s) lower than the target number are delivered to the user. Using a heuristic filter allows many spam filtering methods to be used, resulting in better performance than any single method by itself.

2.3 Study Motivation

Thus, to re-investigate prediction of oil futures price – our statement of problem is thus:

1. The birth and increased adoption of the mobile smartphone alongside its popularity, access ease, mobility, portability – which has consequently, eased the adoption and adaptability of SMS, have consequently, led to the increase of spam and phishing activities.
2. Companies today, are faced with the challenge of dealing with SMS spam. A major issue has been that existing approaches to resolving SMS spam are imported from successful email anti-spam solutions. Thus, are quite unable to effectively and efficiently tackle SMS spam successfully – as their performance is seriously hampered and degraded by the parametric feats used to filter spams.
3. Spams have continued to soar with the advent of SMS. The alarming growth rate of spams with SMS popularity have now created a propitious environ for spammers to exploit subscribers; Thus, causing both financial loss and emotional instability to users, corporate organs and mobile network operator(s).
4. The formulation and design of an effective SMS filter has continued to suffered setback(s) due to the inherent reason that SMS filters by design are not as simple as email filters due to its limited size of 160-characters of 140bytes sized data. These amongst other constraints, continue to create rippled impediment in size of feature to be selected for training and consequently contributing to poor learning and classification of learning algorithm.
5. In addition, SMS are rippled with slangs, abbreviations, symbols and emoticons that inhibit proper classification of words or texts.

To overcome these amongst many other shortfalls inherent in the adoption of email filters as adapted to handling SMS spam successfully, a hybrid filtering techniques that reduces noise in form of slangs, emoticons, abbreviations in SMS as well as expand message size must be employed to enhance adequate classification. Thus, our research goal(s) is to propose a hybrid deep learning neural network model for text normalization and semantic expansion in SMS spam filtering.

3. MATERIALS AND METHODS

3.1 Dataset Used & Proposed Framework

The dataset used for this study is the knowledge discovery in databases (KDD 2022). The schematic diagram of the proposed framework is seen as in figure 1 with the framework grouped into four (4) sections namely (Ojugo & Eboka, 2018, 2019b, 2020b; Okobah & Ojugo, 2018):

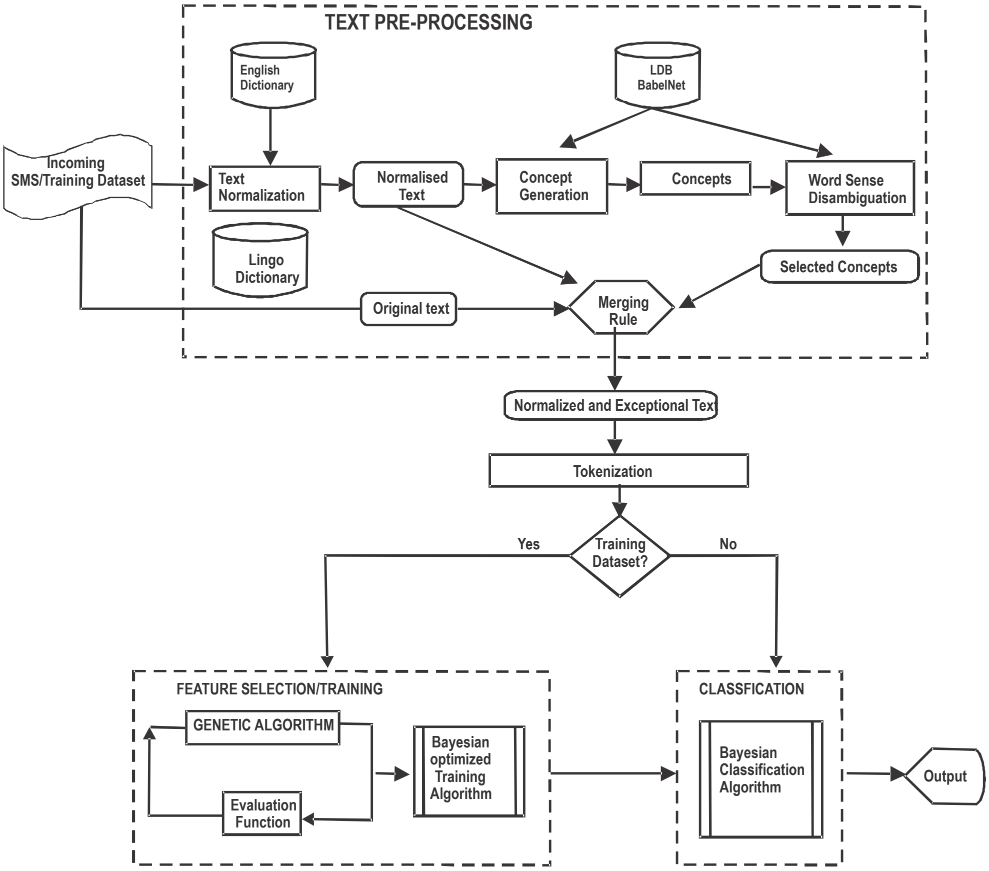


Figure 1: Schematic diagram of the Memetic Algorithm

3.1.1. Pre-processing Section

This section consists of the following:

1. Raw text represents the original text from the sender for normalization and expansion.
2. Text normalization employs two dictionaries. The first is an English dictionary to check if the text are English so as to then normalize text to its root form. The second is a slang dictionary used to translate slang on to English text. The basic operation here, is to replace slang and abbreviation with regular English words from these dictionary. The Freeling English dictionary and No slang dictionary are proposed.
3. Word sense disambiguation: With a variety of concept generated, it seeks to find the concept that is more relevant according to the context of the original message, among all generated concepts related to a certain words. It equally relies on concepts are provided by Language Data Base (LDB) BabelNet repository
4. Tokenisation unit: Tokenization is the process of breaking down a text corpus into individual elements that serve as

input for various natural language processing algorithms. Normalised texts are broken into individual words and stop words and punctuation characters are equally removed in this unit.

5. Merging Rule: Parameters that define the combination of result of pre-processing (original text, normalization and disambiguation stage). Merging rule answers the question from each stage: (a) should it keep the original token(s)?, (b) should text normalization be performed?, (c) should it perform concepts generation?, and (d) should it perform the word sense disambiguation?
6. Concepts generation are semantically analyzed already normalized text to deduce their concept. The concepts are provided by Language Data Base BabelNet repository.
7. Normalized and Expanded text is a combination of text obtained from various output of preferred stages of the pre-processing model.

3.1.2. Feature Selection / Training Section

Need to minimize the number of features as input parameters for classification – since, an increase in the number of features used will add to the computational complexity of the system. Thus, the CGA algorithm is used in selection of feats obtained from pre-processing. Input is the dataset (tokens of normalized/expanded text from the text pre-processing section). The model is made up of the following sections:

1. GA Unit – yields rule-based, genetic rep of normalized and expanded test defined. Algorithm then initializes model with a random population that is created and subjected to repetitive application of recombination, mutation, inversion and selection operators to improve generated population from the original dataset.
2. Evaluation Unit contains fitness function that measures the quality of represented candidate solution(s). It computes the optimality of a solution by comparing a chromosome against all other chromosome using predefined function.
3. Training Unit: Trains the filter based on Bayes Probability using the SMS corpus of spam and genuine texts. A collection of tokens appearing in each corpus and their total occurrences (scores) are maintained in the database so that based on their occurrences, each set of spam and genuine data is assigned a criterion or probability score for its capacity of determining a text or message to either be a spam or genuine text.

3.1.3. Classification Section

Based on the frequency probability of occurrence of each word (tokens) as spam or legitimate, each incoming unseen normalized message data is processed and classified as either legitimate or spam by the Bayesian classifier. In the event of misclassification, users can rectify this classification by reading the message and re-adding the message to inbox. This will automatically correct and update the database for future classification. Thus, making Bayesian filters quite adaptive.

3.1.4. Output Section

Resulting filter, as either a spam or ham – is the expected output of this unit.

3.2. The Experimental Framework

Ojugo and Eboka (2021) and Ojugo and Oyemade (2021) described in details the workings, structure and feature selection for a hybrid genetic algorithm trained Bayesian network (Ojugo & Eboka, 2021; Ojugo & Oyemade, 2021). First, the GA-BN unit is initialized with a set of (n-r!) individual if-then, fuzzy rules. The individual fitness for each of the rule-set is then first computed, for the first pool of 30-individuals that are selected from the first parents using the **tournament** method. This method helps the Cultural GA trained Bayes network to determine the new pool as well as candidate solutions to be selected for mating. With candidate solutions that meets the fitness criteria selected – the model then applies crossover and mutation operators to the pool to effectively help the network to learn the dynamic, chaotic and complex non-linear underlying feats of interest. For this study, we use a multi-point crossover on the pool to yield new parents. The new parents contribute to yield a new pool of individuals. Mutation is reapplied and individuals are allotted new random values that still conform to the belief space. The mutation applied depends on how far CGA is progressed on the net and how fit the fittest individual in the pool (i.e. fitness of the fittest individual divided by 2). New individuals replace old with low fitness so as to create a new pool. Process continues until individual with a fitness value of 0 (i.e. solution) is found (Ojugo & Eboka, 2020b, 2021; Ojugo & Nwankwo, 2021; Ojugo & Otakore, 2021).

Table 1: Rule-Based Encoded Score

Code	Rule Input Parameters	Genuine	Spam
P01	Message Size	0.50	0.50
P02	Message Character	0.50	0.50
P03	Message From	0.50	0.50
P04	Message To	0.50	0.50
P05	Subject	0.30	0.70
P06	Body of Message	0.25	0.75

Fitness function (*f*) is resolved with initial pool (Parents) using the genuine class as thus:

R1:50 R2:50 R3:50
R4:50 R5:30 R6:50

Table 2: 1st/2nd Generation from Parents

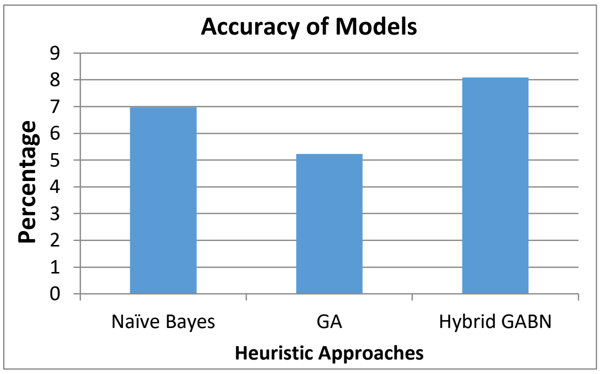
	Selection	Chromosomes (Binary 0 and 1)			Fitness Functio
		1st Gen	Crossover	2nd Gen	
1	50	110010	1 and 6	1100**01**	49
2	50	110010	2 and 5	1100**10**	50
3	50	110010	3 and 6	1100**01**	49
4	50	110010	4 and 5	1100**10**	50
5	30	011110	5 and 6	0111**01**	29
6	25	011001	6 and 5	0110**10**	26

Initialization/selection via Bayesian network ensures that first 3-beliefs of the cultural GA-unit is met; While, the mutation operator for the GA ensures that the fourth belief is met. Its influence function influences how many mutations take place, and the knowledge of solution (how close its solution is) has direct impact on how algorithm is processed. Algorithm stops when best individual has fitness of 0.3 (Ojugo et al., 2021; Ojugo & Eboka, 2020b). Model stops if stop criterion is met. GA-BN utilizes the number of epochs to determine stop criterion.

4. RESULT FINDINGS & DISCUSSION

4.1. Accuracy and Convergence Time

Using naïve Bayes and GA (as standalone model) benchmark to ascertain how well our hybrid GABN algorithm performed, we obtain the results in fig 2 and fig 3 respectively. It shows that hybrid GABN outperforms Naïve Bayes and GA models. However, for the mean processing time required to converge – it is found that GABN performed least. This can be attributed to the fact that: (a) the hybrid model needs to first use GA as pre-processor to train Bayesian network, (b) for such hybrids, there are always structural dependencies with the underlying heuristics employed/merged and conflicts in data encoding that is required. These must be resolved in order for the model to perform appropriately.



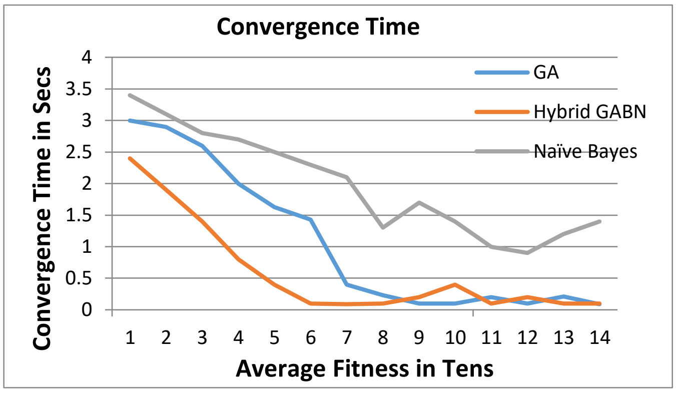


Figure 2: Comparative Accuracy of models

4.2 Model Evaluation/Findings Discussion

Here, to compute accuracy, recall, error rate (ER) and specificity are used to evaluate the performance of the detection models (Brofman Epelbaum & Garcia Martinez, 2014; Gokarn & Choudhary, 2021; Zawislak et al., 2022). The formulas of the above criteria are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Error\ Rate = \frac{FP + FN}{FP + TP + TN + FN} \quad (3)$$

$$SPEC = \frac{TN}{TN + FP} \quad (4)$$

A true positive (TP) is a case (rule) that correctly distinguishes spam from genuine text. The true negative (TN) shows normal text message data classified correctly as normal. The false negative (FN) denotes a case in which a text is classified as normal data, and a false positive (FP) means that a normal text is classified as a spam. The accuracy rate shows the overall correct detection accuracy of the dataset, ER refers to the robustness of the classifier, recall indicates the degree of correctly detected attack types of all cases classified as attacks, and specificity shows the percentage of correctly classified normal data. In the above, higher accuracy and recall and lower ER indicate good performance.

To further measure effectiveness and accuracy, we measure their rate of misclassification and corresponding improvement percentages in both training and test data sets as summarized in Tables 3 and 4 respectively. Equations for misclassification rate and its improvement percentage of unsupervised (B) model against supervised (A) model respectively, is calculated as follows:

Figure 3: Comparative convergence Time

$$Misclassification\ Rate = \frac{No.of\ Incorrectly\ Classified\ Rules}{No.of\ Sample\ set} \quad (5)$$

Table 3: Misclassification Rate of Each model

Model	Classification Errors	
	Training Data	*Testing Data*
Naïve Bayes	52.5%	23.2%
Genetic Algorithm	48.4%	4.7%
GABN	19.6%	1.02%

Also, its improvement percentage is computed as thus:

$$Improvement = \frac{MR(A) - MR(B)}{MR(A)}\ x\ 100 \quad (6)$$

Table 4: Improvement Percentage

Model	Improvement %	
	Training Data	*Testing Data*
Naïve Bayes	2.11%	3.6%
Genetic Algorithm	2.32%	4.02%
GABN	0.09%	0.12%

Tables 3 and 4 respectively shows misclassification error rate with Naïve Bayes, GA and GABN at 23.2%, 4.7% and 1.02% (i.e. error rate in false-positive and true-negative) respectively; Consequently, they all promise an improvement rate as of 3.6%, 4.02% and 0.12% respectively.

5. CONCLUSION

SMS spam filters can have the capacity, and be granted capability to transcribe slangs, emoticons and abbreviations into standard terms as well as expand the message size to enhance better feature extraction for classification algorithms and approaches. The study will also serve to reduce orthographic error found in SMS, chat groups and other social network communication medium that impedes machine learning algorithm. This is because from the various approaches adopted to SMS spam filters – the content-based models with text pre-processing has shown to perform better.

Machine translation performs better when applied to normalized text messages. We can equally combined multiple approaches in noisy data, text normalization to create a better output. But, extracting only relevant feats to train the classifier has been reported to contribute to the efficiency of SMS spam filters. Thus, we propose text preprocessing SMS spam filter model with the capability of normalizing, expanding text messages and extracting suitable features as dataset input parameters for training the adopted classification algorithm and model.

Conflict of Interest

The authors declare that there is no conflict of interest.

REFERENCES

Akazue, M. I., Ojugo, A. A., Yoro, R. E., Malasowe, B. O., & Nwankwo, O. (2022). Empirical evidence of phishing menace among undergraduate smartphone users in selected universities in Nigeria. *Indonesian Journal of Electrical Engineering and Computer Science*, *28*(3), 1756. https://doi.org/10.11591/ijeecs.v28.i3.pp1756-1765

Amalraj, J. R., & Lourdusamy, R. (2022). A Novel Distributed Token-Based Access Control Algorithm Using A Secret Sharing Scheme for Secure Data Access Control. *International Journal of Computer Networks and Applications*, *9*(4), 374. https://doi.org/10.22247/ijcna/2022/214501

Br, M., Haider, P., & Scheffer, T. (2021). Highly Scalable Discriminative Spam Filtering Challenges in Email Classification. *National Cybersecurity Institute*, *34*, 1–11.

Brofman Epelbaum, F. M., & Garcia Martinez, M. (2014). The technological evolution of food traceability systems and their impact on firm sustainable performance: A RBV approach. *International Journal of Production Economics*, *150*, 215–224. https://doi.org/10.1016/j.ijpe.2014.01.007

Ezpeleta, E., de Mendizabal, I. V., Gómez Hidalgo, J. M., & Zurutuza, U. (2020). Novel email spam detection method using sentiment analysis and personality recognition. *Logic Journal of the IGPL*, *28*(1), 83–94. https://doi.org/10.1093/jigpal/jzz073

Ezpeleta, E., Zurutuza, U., & Gómez Hidalgo, J. M. (2016). *Does Sentiment Analysis Help in Bayesian Spam Filtering?* (pp. 79–90). https://doi.org/10.1007/978-3-319-32034-2_7

Gokarn, S., & Choudhary, A. (2021). Modeling the key factors influencing the reduction of food loss and waste in fresh produce supply chains. *Journal of Environmental Management*, *294*, 113063. https://doi.org/10.1016/j.jenvman.2021.113063

Hayati, P., Potdar, V., Talevski, A., Firoozeh, N., Sarenche, S., & Yeganeh, E. A. (2010). Definition of Spam 2.0: New spamming boom. *4th IEEE International Conference on Digital Ecosystems and Technologies - Conference Proceedings of IEEE-DEST 2010, DEST 2010*, *May 2014*, 580–584. https://doi.org/10.1109/DEST.2010.5610590

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., & Alegre, E. (2022). A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10195-4

Jáñez-Martino, F., Fidalgo, E., González-Martínez, S., & Velasco-Mata, J. (2020). Classification of Spam Emails through Hierarchical Clustering and Supervised Learning. *National Cybersecurity Institute*, *24*, 1–4. http://arxiv.org/abs/2005.08773

Kamoru, B. A., Jaafar, A. Bin, Murad, M. A. A., Ernest, E. O., & Jabar, M. B. A. (2017). Spam Detection Approaches and Strategies: A Phenomenons. *International Journal of Applied Information Systems*, *12*(9), 13–18. https://doi.org/10.5120/ijais2017451728

Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. *IEEE Access*, *7*, 168261–168295. https://doi.org/10.1109/ACCESS.2019.2954791

Karimi, Z., Mansour Riahi Kashani, M., & Harounabadi, A. (2013). Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods. *International Journal of Computer Applications*, *78*(4), 21–27. https://doi.org/10.5120/13478-1164

Lazarevic, A., Kumar, V., & Srivastava, J. (2005). Intrusion Detection: A Survey. In *Managing Cyber Threats* (Issue January). https://doi.org/10.1007/0-387-24230-9_2

Nivedha, M. A., & Raja, S. (2022). Detection of Email Spam using Natural Language Processing Based Random Forest Approach. *International Journal of Computer Science and Mobile Computing*, *11*(2), 7–22. https://doi.org/10.47760/ijcsmc.2022.v11i02.002

Ojugo, A. A., & Eboka, A. O. (2018). Comparative Evaluation for High Intelligent Performance Adaptive Model for Spam Phishing Detection. *Digital Technologies*, *3*(1), 9–15. https://doi.org/10.12691/dt-3-1-2

Ojugo, A. A., & Eboka, A. O. (2019a). Extending

Campus Network Via Intranet and IP-Telephony For Better Performance and Service Delivery: Meeting Organizational Goals. *Journal of Applied Science, Engineering, Technology, and Education*, *1*(2), 94–104. https://doi.org/10.35877/454ri.asci12100

Ojugo, A. A., & Eboka, A. O. (2019b). Signature-Based Malware Detection Using Approximate Boyer Moore String Matching Algorithm. *International Journal of Mathematical Sciences and Computing*, *5*(3), 49–62. https://doi.org/10.5815/ijmsc.2019.03.05

Ojugo, A. A., & Eboka, A. O. (2020a). An Empirical Evaluation On Comparative Machine Learning Techniques For Detection of The Distributed Denial of Service (DDoS) Attacks. *Journal of Applied Science, Engineering, Technology, and Education*, *2*(1), 18–27. https://doi.org/10.35877/454ri.asci2192

Ojugo, A. A., & Eboka, A. O. (2020b). Memetic algorithm for short messaging service spam filter using text normalization and semantic approach. *International Journal of Informatics and Communication Technology (IJ-ICT)*, *9*(1), 9. https://doi.org/10.11591/ijict.v9i1.pp9-18

Ojugo, A. A., & Eboka, A. O. (2021). Empirical Bayesian network to improve service delivery and performance dependability on a campus network. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *10*(3), 623. https://doi.org/10.11591/ijai.v10.i3.pp623-635

Ojugo, A. A., & Nwankwo, O. (2021). Multi-Agent Bayesian Framework For Parametric Selection In The Detection And Diagnosis of Tuberculosis Contagion In Nigeria. *JINAV: Journal of Information and Visualization*, *2*(2), 69–76. https://doi.org/10.35877/454RI.jinav375

Ojugo, A. A., Obruche, C. O., & Eboka, A. O. (2021). Quest For Convergence Solution Using Hybrid Genetic Algorithm Trained Neural Network Model For Metamorphic Malware Detection. *ARRUS Journal of Engineering and Technology*, *2*(1), 12–23. https://doi.org/10.35877/jetech613

Ojugo, A. A., & Otakore, O. D. (2021). Forging An Optimized Bayesian Network Model With Selected Parameters For Detection of The Coronavirus In Delta State of Nigeria. *Journal of Applied Science, Engineering, Technology, and Education*, *3*(1), 37–45. https://doi.org/10.35877/454RI.asci2163

Ojugo, A. A., & Oyemade, D. A. (2021). Boyer moore string-match framework for a hybrid short message service spam filtering technique. *IAES International Journal of Artificial Intelligence*, *10*(3), 519–527. https://doi.org/10.11591/ijai.v10.i3.pp519-527

Okobah, I. P., & Ojugo, A. A. (2018). Evolutionary Memetic Models for Malware Intrusion Detection: A Comparative Quest for Computational Solution and Convergence. *International Journal of Computer Applications*, *179*(39), 34–43. https://doi.org/10.5120/ijca2018916586

Rathi, M., & Pareek, V. (2013). Spam Mail Detection through Data Mining – A Comparative Performance Analysis. *International Journal of Modern Education and Computer Science*, *5*(12), 31–39. https://doi.org/10.5815/ijmecs.2013.12.05

Reads, C. (2014). *An Introduction to Intrusion-Detection Systems An Introduction to Intrusion-Detection Systems. January 2009*.

Redondo-Gutierrez, L. Á., Jáñez-Martino, F., Fidalgo, E., Alegre, E., González-Castro, V., & Alaiz-Rodríguez, R. (2022). Detecting malware using text documents extracted from spam email through machine learning. *Proceedings of the 22nd ACM Symposium on Document Engineering*, 1–4. https://doi.org/10.1145/3558100.3563854

Sahmoud, T., & Mikki, D. M. (2022). *Spam Detection Using BERT*. https://doi.org/10.48550/arXiv.2206.02443

Shelke, Y., & Sharma, A. (2020). Internet of Medical Things. *Technology Intelligence and IP Report: Thematic Report*, *28*, 2–30.

Yao, J., Wang, C., Hu, C., & Huang, X. (2022). Chinese Spam Detection Using a Hybrid BiGRU-CNN Network with Joint Textual and Phonetic Embedding. *Electronics*, *11*(15), 2418. https://doi.org/10.3390/electronics11152418

Yu, Y., Li, M., Liu, L., Li, Y., & Wang, J. (2019). Clinical big data and deep learning: Applications, challenges, and future outlooks. *Big Data Mining and Analytics*, *2*(4), 288–305. https://doi.org/10.26599/BDMA.2019.9020007

Zawislak, P. A., Reichert, F. M., Barbieux, D., Avila, A. M. S., & Pufal, N. (2022). The dynamic chain of innovation: bounded capabilities and complementarity in agribusiness. *Journal of Agribusiness in Developing and Emerging Economies*. https://doi.org/10.1108/JADEE-04-2021-0096