Nat. Volatiles & Essent. Oils, 2021; 8(5): 621 – 633

NVEO
Natural Volatiles & Essential Oils

A Gwhsc - A Genetic Algorithm Based Weighted Hybrid Classifier For Sms Spam

Dr.K.S.Bhuvaneshwari[1], Dr.M.Vigenesh[1], Dr. D. Bhanu [2], Dr.S. Kannimuthu[3]

[1]Karpagam College of Engineering, Coimbatore, Tamilnadu, India -641032
[1]Karpagam Academy of Higher Education, Coimbatore, Tamilnadu, India -641032
[2]Karpagam Institute of Technology, Coimbatore, Tamilnadu, India -641105
[3]Karpagam College of Engineering, Coimbatore, Tamilnadu, India -641032
E-mail: bhuvana.me@gmail.com[1], vigenesh.murugesan@kahedu.edu.in[1] bhanu.saran@gmail.com[2,] kannimuthu.me@gmail.com[3]*

Abstract:

The proliferation in the volume of unwanted SMS called spam has paved strong need for the development of more reliable and robust anti-spam detectors. In recent years, machine learning algorithms are being successfully used for classifying the datasets which are successfully adopted in many applications. Machine learning methods are also can be applied to detect the SMS spams. Feature selection in Machine Learning process plays a vital role in improving the accuracy and other performance aspects. In this work, a novel GWHSC approach for detecting SMS spam is proposed that utilizes evolutionary method for feature selection which is realized to be the efficient method for spam filtering. Experimental result show that proposed work outperforms well when compared to the existing approaches.

Keywords:- Genetic Algorithm (GA), Evolutionary algorithms, Feature Selection, Machine Learning, SMS Spam, Spam Filtering.

1.0 INTRODUCTION

In recent world with the modern advancements in mobile technology, Short Message Service (SMS) is considered as one of the most important and prevalent communication amenities for millions of people over mobile devices. SMS service gains its popularity for its easiness, simple to use and it's less cost as it is available for the layman. Not only for the end-users, but also for the service providers ROI (return on investment) is good enough. SMS is extensively involved in various applications including mobile banking, insurance sectors, automatic information retrieval, recent updates, and advertisements for discounts and for promoting the introduction of various new products, web alerts for user authentication.

Thus, sending of SMS over network increases two times every decade which increases the problem of spams. Spam a short message is termed as an unsolicited annoying bulk messages sent over the internet becomes a big problem in recent days. Mostly these unwanted bulk messages are commercial in nature and the number varies from place to place and also over time it varies. Lot of people including the service provider and the government are involved in spam control process and uses many measures. These measures sometimes reduce the junk and spam messages in number but not completely as the initiator cannot be identified properly. Spam messages are really a threat to the internet users as it utilizes high bandwidth, wastage of memory and processing time. It prevents the mobile users from making a quality use of time, memory utilization which shows a destructive consequence and finally a drift in the performance of the entire communication system. The mobile users who receive these unintentional spam messages irritates them and spoils their mental health.

These spams also make the mobile users as victims for certain financial hindrances caused by fraudulent events of spammers, who send fake messages pretending to be from some reputed financial organizations like banks, insurance sectors with the bad intention to persuade individuals to disclose sensitive personal details such as mail passwords, account number, pin number ect. Spam is mushroomed as one of the widespread threatening issue that influences the society and in turn the mobile users. Hence it is a high demanding need for designing an automatic filtering system for SMS spams. These SMS is a short message whose length is limited to less number of bytes which is around 140. These short messages contain informal phrases and acronyms making it more difficult for identifying spams from it. Another factor which is predominantly makes the spam detection more difficult is the missing of priority over the SMS messages. The system could not able to maintain the hierarchy over the messages as which is more relevant and needed.

As the SMS Spams are considered to be a high-risk factor to the mobile users, many techniques were being applied by the service providers for sensing and segregating spam over required messages. Because of its increased attention, SMS spam detection has been one of the recent research issues. Hence, a pragmatic study to analyze the use of various evolutionary classification algorithms to deal with this serious threat of detecting spam messages and a new framework is being proposed. The contributions of this paper are narrated as follows:

- A hybrid genetic algorithm based technique experimented over SMS corpus for classifying spam and ham messages with higher accuracy.
- A detailed analysis of various machine learning algorithms also done for a comparative study with proposed one.
- A proof of concept also narrated showing a SMS messages efficiently classified.

The paper is organized as given below. In section 2, a deep study over the existing methods of spam detection is performed. Section 3 focusses on the bird's view of SMS spam detection and filtering techniques. Prevailing machine learning algorithms for spam filtering are discussed. In section 4, the proposed framework for spam detection is experimented and results were discussed. A detailed comparative analysis is done performed with the existing methods and the proposed one. For comparing the machine learning algorithms, various standard measures are employed to evaluate the strength of it. Out of many such standards, here we use Sensitivity, Specificity, ROC curve, Accuracy and Correctness for measuring their efficiency. Finally, the conclusion for the paper is detailed.

2.0 RELATED WORKS

The spam filtering technique proposed in [1] is a two-fold approach providing spam detection accurately at byte-level. The methodology during its training phase receives the messages labelled with spam from the inbox and send it to the next phase called spam detection system. The classifier used in the framework, learns the rules based on different encoding techniques advanced based on the byte-level transition of spam labelled messages. The main advantage is that the classifier is immediately updated over newer mobile devices when there is change over in the spam messages.

In paper [2], mobile service providers, cellular device operators formulated a well-defined method to defend the users of mobile devices from unintentional SMS spam messages using Open Mobile Alliance (OMA). The main advantage of this method is it works well for many of the mobile users to detect the process from the administrative level. Many spams are detected at the administrative level but it does not provide solution for customized filtering capability. In paper [3][4], a content-based spam filtering method is adopted which makes use of direct adaptation of feature specific terms, bi-grams and tri-grams for filtering the spam from ham messages.

In paper [5], an efficient bi-level text classification scheme is proposed for SMS spam filtering and segregate the messages into prioritized and non-prioritized messages. In level 1, a binary two-level classifier is involved in dividing the messages into spam and non-spam classes. Subsequently in level 2, the non-spam messages are arranged in a hierarchy based on their priority. In this work, popular classifiers such as Naïve Bayes, Support Vector Machine (SVM), Latent Dirichlet Allocation (LDA) and NMF Non-Negative Matrix Factorization are used for organizing the SMS at diverse levels. In paper [6] Gomez et al. revealed the application of statistics-based learning methodology in which the classifier trains the lexical word by word features including tri-grams and n-grams for dropping out the spam messages. It tests the applicability of Bayesian classification model for evaluating the filtering of spams. They have specifically tested the feasibility of applying Bayesian based classifier to SMS spam problem. This paper also provides a deep insight into the application of SVM classifier model for SMS and also email detection.

This paper [7] addresses the usage of content-based learning process for spam detection. The authors make use of the short informal regional words, idioms and phrases, short acronyms for its processing. SMS Assassin a model is been introduced for spam SMS filtering which incorporates the SVM technique which is proved to be one of the effective ML based learning mechanism. Researches used crowd sourcing for keeping it more updated as the phrases and idioms used by the users keep changing over time and the result evaluated is more interesting and efficient. This technique though it provides high accuracy at client-side filtering but fails to handle at the server side. In paper [8], security threats due to spam SMS and the problems arise are discussed. The reason behind the adaptation of machine learning to the problem of spam detection is vibrant as the spammers keep on using innovative techniques for spamming. So, a technology which handles real world, reliable, platform independent, highly massive which is none other than machine learning is been used. ML algorithms are utilized because of its capability to handle huge volume of data with varied dimensions at a particular time. So there exist a need for implementing an adaptable intelligent spam filtration model at the user side based on amalgamation of two methods Genetic Algorithm (GA) and RWN is discussed by the author. Sequence of experiments are carried out intensively based on SMS corpus. The results prove that there is a betterment in the proposed algorithm in terms of precision, accuracy and recall value in addition to automatic filtering of appropriate features.

In paper [9], Metsis et al. variety of Naïve Bayesian (NB) Models including Bernoulli Model, Multinomial Method NB, Gaussian Naïve Bayes Model and flexible NB Mode. A comparative study over all these NB models is performed and the result shows that the Multinomial and flexible makes a better accuracy. In [10] Sahami et al. used normal Bayesian model for detecting and classifying junk messages. In paper [11,12] utilizes ANN artificial neural network base spam detection methodology,

whereas [13] makes use of kNN and [14] shows the applicability of Decision Tree for spam filtration and detection process.

In paper [15], Uysal et al. analyzed the effectiveness of many feature filtration and selection mechanisms in different colloquial languages like Turkish and English. This paper makes use of text patterns or texts as key for filtering process. Feature extraction process selects many text messages from the corpus containing spams and ham messages. It consists of BOW bag-of-words method and second uses structural approach for dealing with spams. BoW ha many features to be extracted during feature filtration. Many usual patterns input exists among the SMS were learnt during model building phase and is used in classification of spam and ham. This hybrid approach of utilizing BoW and structural patterns shows their improvements during experimentation. In paper [16], Ahmed et al. recommends a dual approach for differentiating ham and spam during SMS segregation by using Apriori and NB Naïve Bayes approach for text classification. The state of art algorithm boosting technique is presented by Akbari et al. Stem Words were involved in feature selection which removes unexploited features without relaxing the accuracy. Some practitioners make use of deep neural nets RBM restricted Boltzman Machine for SMS classification into ham and spam over UCI spams datasets. This proposed algorithm provides an accuracy of 98% when it is compared with other basic machine learning algorithms such as NB, random forest and SVM.

Based on the literature study, some of the seed parameters required for developing a model for SMS spam detection and filtering is summarized below. As the growth of the SMS users is increasing tremendously, the number of spammers is also increasing drastically which acts as a base for designing a highly flexible algorithm that handles real time spam detection that makes the user feel as if there is no spam filtering mechanism used. The algorithm designed for this should be less complex, more efficient in terms of processing with high accuracy since it is embedded in the mobile device itself. The algorithm should react to user preferences and provide with personalized spam detection and filtering as it varies from person to person. The algorithm should be designed to handle both servers as well in the client side. The service providers and the governments must provide some sort of methodologies to reduce number of spams and also at the client end user mobile devices too. The algorithm or the method proposed should ensure the security of applying this at the client side. The methodology adopted should accustom any kind of devices and works in any platform providing platform independence property.

3.0 ML based Classification Models for Spam Detection

In the following, we investigate some well-known classification methods employed in spam filtering applications. To name a few classical Naïve Bayes, Genetic algorithm, Lazy K-Star. K-Nearest Neighbours (KNN), Voted Perceptron and SVM.

1) Naive Bayes Algorithm (NB): A probabilistic methodology, Naïve Learning Bayes (NB) algorithm is a very basic and most important method for e-mail and SMS spam learning. The base for this algorithm is independent hypothesis based on Bayes Theorem which makes use of probabilities and produce a statistical classification algorithm. This property makes the NB to treat each and every parameter that defines the item is independent of each other. The algorithm shows a betterment in efficiency, though the hypothesis used is not true for all

situations. As the variable or the attributes are seeming to be independent, the variance for each class label only used for classification. So, this can be utilized as an introductory model for learning the basic of how spam leaning is done. This model can be used to find out the probability of belongingness of a particular object to its intended class or a category.

2) C4.5 Algorithm: A conventional learning methodology C4.5 is a decision tree-based algorithm utilized for enhanced character-based classification problem. Because of its standard, simple, consistent and exhaustive architect it's been widely utilized.
3) Support Vector Machine (SVM): Support Vector Machine (SVM) a supervised classifier, classifies the objects based on the hyper planes that performs a boundary-based classification to split the class labels and also for regression synthesis and analysis. The new datasets are classified based on these gaps. In the hyperplane the example data is represented as individual points in space where these data are separated by clean gap for segregating into different categories. Higher dimensional non-linear mapping is used to transform training data. The model looks for a linear hyper plane within this higher dimension obtained from the training set. With the advent of these higher dimensional linear hyper planes, classification object belonging to different class is done using margins. The method builds a classifier based on the training datasets taken as an example which belongs to one or more category, classifies the new object using the model built. Its main advantage is its limited usage of training samples, which uses only the objects available nearer to the border for classification. The appropriate utilization of kernel seed functions is employed in polynomial classifier building, radial base function and neural networks for universal learning capability.
4) Ada Boost: Robert Schapire and Yoav Freund found the adaptive boost technique in short termed as AdaBoost, an automated meta algorithm in machine learning works based on rule generation. This were utilized by many researchers in particular, Zhang et al. [17] used as one of the machines learning classifying technique for filtering out the spams at china. It uses many types of filters base on content, structural etc., It presented the outcome of the algorithm using ROC which filter outs the spams more efficiently. Boosting combines many fairly strong, weak and also inaccurate set of rules and works in a repeated learning mode for improving the learning capability. The weak rules involved in classification can be improved by assigning a weight and is turned into stronger.

3.1 Limitations of Traditional Non-Automated Spam Filtration Techniques

In this evolutionary real world, the usage of SMS is enormous which increases the number of spammers. Lot of mechanisms were used to overcome the issues generated by them. Many automated, semi-automated and non-automated mechanism were utilized to detect and filter out the spams and ham messages. It is not practically feasible to build and utilize non-automated spam filtration as it has lot of disadvantages. This section briefs about the limitations of non-automated spam filtering techniques and need for automated machine learning algorithms for spam detection and filtering. Many semi-automated frameworks were into practice but didn't solved the problem completely as it possess its own bottlenecks in identifying spam patterns and its structure. Simple and foremost technique for filtering is by making use of certain pattern in the text messages making it to separate into spam or ham messages.

Regular patterns are identified as regular expressions are involved in finding the frequently occurring patterns making it easy to segregate spam and ham message. But this method fails in this real time implementation of it where the spammers identifies the rules for pattern definition and hacks the filtration process and violates it. Many of the users makes use of a traditional way of handling spams just by blocking the sender who sends the unwanted messages not intended for the user. But the drawback in this is the spammers keeps different sending addresses for broadcasting their spams and makes the entire system tiresome in updating the database. This is one of the proven concepts that it is not efficient in this practical world where the spammers mushrooming every moment. Many practitioners tried utilizing certain keywords and its matching process, area wise filtering of messages, block-listing and grey-listing techniques. These techniques where been exploited by the spam generators effortlessly. Most of these techniques pull downs the performance and increases true negatives and false positives during classification process.

4.0 DATASET COLLECTION

At the outset, set of sample SMS are collected and are processed for outliers to get a structured class labels and messages. Then, these samples are transformed by utilizing text embedding techniques such as Filter Stop words, Stemmers such as Poster, Xerox, Dawson, Lovins and N-Gram. Tokenizing, and filter words by length. The dataset used in our study is a huge SMS Spam Corpus filtered out with only two columns one containing SMS text messages and the other consisting of class label which are filtered out. To filter out the text messages, the machine learning based algorithm proposed proved to be an effective technique to classify and filter stem words. A sample of around 2k data is been taken from the dataset containing both spam and ham messages which are scrutinized extensively for filtering out required texts. These sample text messages gathered from the dataset are undergone experiments with known machine learning algorithms and their performance are noted and tabulated for metrics such as root mean square, recall value, precision, accuracy and error. The model is tested with usual classifier algorithms such as random forest, decision tree and SVM.

5.0 THE BENEFITS OF OUR PROPOSED ML BASED APPROACH

Automated technique for spam filtering is quite a reasonable attempt in this emerging world. Machine learning with deep and reinforcement learning capabilities is proved to be better in recent times. Adaptive machine learning algorithms are designed to accustom to the changing patterns of spam generators. Spammers keep upgrading the newest methods for scattering spam SMS and e-mails. Dynamic filters are indeed required for handling these subtle deviations where static filtering techniques fails in it. With the advent use of two categories of machine learning, both supervised (most widely utilized) and unsupervised is into practice. Supervised learning methodology applies class labels for classifying new training sample data for classifying it. Supervised methods generate already a model with the known labels performs better with the available set of more training data.

With this brief discussion, a novel hybrid supervised classifier with automated learning property is needed by the current messaging framework. The algorithm is termed as **GWHSC-a genetic based weighted hybrid SMS classification algorithm for** e-mail and SMS spam detection with greater accuracy and confidence. The algorithm finds out whether the message or e-mail is spam or ham by comparing it with the available database of black-listings and the stem words from the message is

compared with collection of spam structured text patterns. The experimental results shown that the fusion technique proposed have a substantial enhancement in spam filtering and detection process.

6.0 METHODS OF RESEARCH

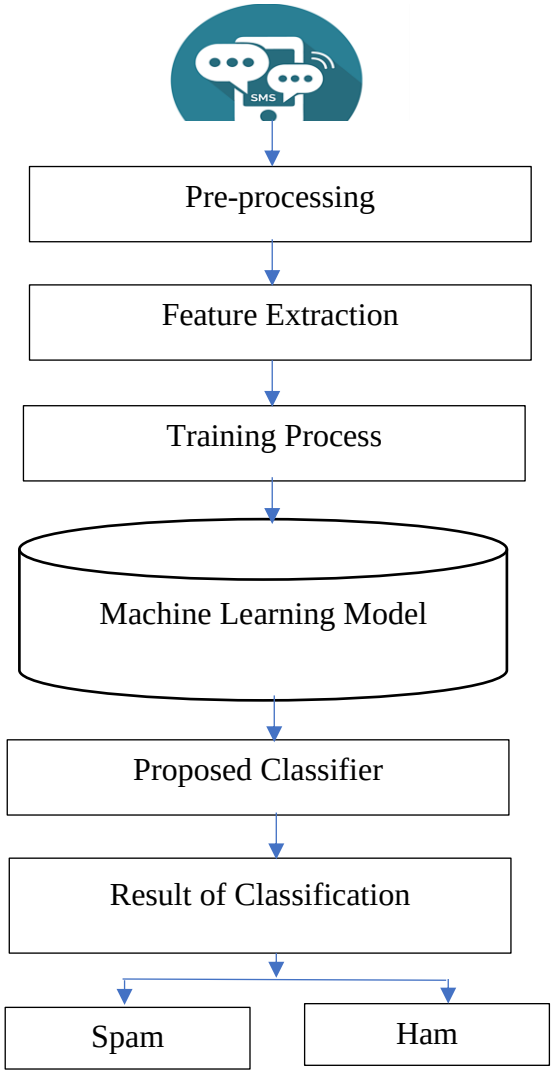

Figure. 1. Architecture of SMS Spam Classifier

6.1 Evolutionary Algorithms

An adaptive learning algorithm for many real-life applications using some standard rules are employed with evolutionary regulations. Some kind of probabilistic search techniques and also optimization heuristic are utilized for finding solutions that makes use of simulation of biological genes and in each generation with more populations are evolved over time using standard Darwin theory. It makes use of selection process, followed by recombining and in-turn mutation for evolving better process and a known methodology in this arena are genetic algorithms which involves more randomized and robust search.

6.2 **Steps involved in SMS spam proposed algorithm:**

Figure 2 presents the methodology of the proposed work. The work is done in steps in which the first would be SMS dataset collection. The collected SMS spam data is retrieved for pre-processing analysis. During the second phase, from the pre-processed dataset appropriate features describing the spam and ham SMS are separated by applying evolutionary hybrid genetic algorithm-based optimization which further improves the classifier performance. Final phase of the spam detection is done with applying the logistic regression which is proved to be one of the best classifiers for spam detection.

Genetic algorithm is termed as an evolutionary procedure in which a greater population of varied solutions progresses over a set of newer generations. Succeeding generations recreate fitness of each proven solution for selecting the improved results from the earlier ones based on their fitness standards for reproduction. Survival of fittest principle is used for the segregation of better solutions from the improper one. From the fitness value the "goodness" is defined for every solution. The nominated solutions experience recombining based on the crossover and mutation procedure and occasionally flip operators. From the usual form, the genetic representation of the parameters of the resolutions fluctuates significantly. A variable one, fixed and binary encoded strings becomes the major factor for many researches in genetic domain as they are provided with a greater number of schemata but with simple implementations.

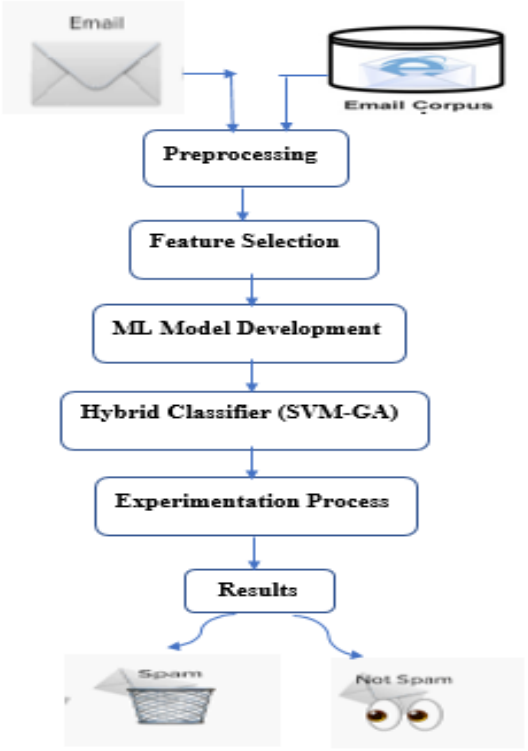

Figure. 2. Architecture of Proposed Hybrid SMS Spam Classifier

6.3 GENETIC ALGORITHM DISSECTION

The supremacy of the genetic algorithm lies within the core concept 'Crossover'. It involves in a

traditional way of systematic conversation of genetic resources amongst solutions that transforms bad solutions into a good one and also good solutions into better. During mutation process, a mutation probability value say 'P' is assigned for every gene pattern of a solution. The uncharted solutions or lost gene materials one in genetic algorithm is been recovered with the help of a key factor called mutation into a partial optimal solution which improves the performance. In order to prevent the hasty convergence of the population into partial optimal solutions, Scaling process is involved in rearrangement of fitness values of some solutions to tolerate a stable discriminating pressure in the population. In unimodal optimization phase, GA combines to smaller generation as possible but in multimodal the GA identifies the area where a global optimal exists which converges to an ideal solution. GA possess hill-climbing property which itself becomes a great disadvantage as it stuck at a local optimal point that becomes vulnerable.

6.4 Crossover variation:

In GA, setting up of the value for 'p' is a crucial factor that decides the performance of it. It acts like a pure random search algorithm, if we choose the 'p' value as a higher value. To maintain the performance in between the mutation is required for preventing from early convergence. The normal range for 'p' is set to be within 0.5 and 1. The crossover probabilities 'p' is a vital parameter which controls the entire operation of the GA. Higher convergence of 'p' leads to faster introduction of recent solutions in the population. When the 'p' increases rapidly, solutions are rapid to selection. The typical values of probability 'p' lies in the range between 0.6 and 1. To refrain the genetic factors mutation is used as a sub-operator. Mutation probability helps in measuring the likeliness of the random variables of the chromosome that is to be flipped into other ones. For example, if the encoding of chromosome is done using a string of binary of length between 1 and 100 with p% mutation probability, it infers that one out of hundred bits is randomly taken out and is been flipped out. GA is applied well for many real-life applications especially in spam SMS and e-mail classification filtering applications. Such group of SMS and e-mails are termed as corpus. These are represented as a chromosome classes which experience genetic crossover operation, mutation process and also the fitness one. Then the steps to be followed for SMS classification into spam and ham is defined using a hybrid GA based algorithm.

6.5 **A (GWHSC) GA- Based Weighted Hybrid SMS Classifier Algorithm Rules for SMS Classification:**

A weighted approach is used for classifying spam and ham from the corpus of SMS. A known weight is assigned for the stem words of "gene" in the test bud and the weight of those from "gene" of spam SMS prototypes are compared for determining the matched gene. If the matched ones are of more percentage say 'p' then it is classified as spam SMS.

Fitness Function:

$$F=\begin{cases}1 & Spam\ SMS\\ 0 & Ham\ SMS\end{cases}$$

The core part of the algorithm lies in selecting the fitness function for segregating spam from the corpus of SMS, as it is dependent completely on the problem itself and hence it can be stated at the

beginning of filtering itself. Over applying a sequence of experiments, fitness function is evolved with a minimum for the existing corpus of 5575 SMS messages.

Fitness Function

$$F=\begin{cases}1 & \text{fitness fn} \geq 2\\ 0 & \text{fitness fn} \leq 2\end{cases}$$

An optimized 2-step hybrid procedure has been developed for selecting features from the corpus of SMS. An optimal set of features are found using genetic algorithm to upsurge the precision of the classifier algorithm. In accordance with this technique, an initial set of population which is randomly selected is used. The genetic algorithm is an adaptive and efficient algorithm uses proportion standards for evaluating the population and the random search is completely dependent on probabilistic and also deterministic rules and regulations. GA is used for its high convergence speed but the better optimization over the results are obtained only when the parameters are selected in a proper way. GA uses Darwinian selection, mutation and change of population applying random search technique for feature selection. Based on the metric, selection of convenient feature is evaluated and selected. Fitness function is reorganized as the optimal set of features are evolved. The algorithmic process is repeated with the number of gene patterns and if it more than it is changed based on generations and the gene pattern is randomly generated. This continues, as long as the optimal subset of features reached. Then the repetition of process is stopped and outputs the optimal feature set.

The following algorithm defines the steps involved in evolving an optimal subset of features from the SMS corpus.

1. Start with the available set of features
2. Initial GA parameters are defined with initial values
3. Initial Population corresponding to size of chromosomes P with f1,f2,f3…fm is defined in accordance with Si vector size.
4. A subset {Si} is chosen based on the available count of features
5. Fitness assessment is done using chosen technique (k-NN)
6. Classification of test data {Si} with the vector selected is done.
7. Then the performance metrics are evaluated and also the fitness quantity for all chromosomes available as a final vector.
8. Chromosomes of higher probability is been selected and new population is created and recombination and mutation is applied for it.
9. Repeat the steps until the convergence of chromosomes are done so that the probability differences between fitness function and Ɛ is lesser.
10. Check for the set of all features selected, it will be finished.

7.0 EXPERIMENTAL RESULTS

Experimentation process is carried over the existing machine learning algorithms such as SVM, Logistic Regression, C4.5, Naïve Bayseian Method and Random Forests and is compared with the proposed hybrid SVM-GA algorithm. Parameters such as accuracy, recall value, error and RMSE are analysed. As it is studied clearly from the Table and Figures below, the proposed algorithm outperforms the other

algorithm by reducing false negatives and false positives and improvement in TP and TN. The accuracy of GA-based algorithm is 94.2% which is higher when compared to other algorithms. Precision, recall value and error for SVM and GA also is convincing when compared to existing algorithms. Figure 3 to Figure 7 shows the plot of performance comparisons various machine learning algorithms with SVM-GA on parameters such as accuracy, classifier error, weighted mean precision, weighted mean recall and RMSE.

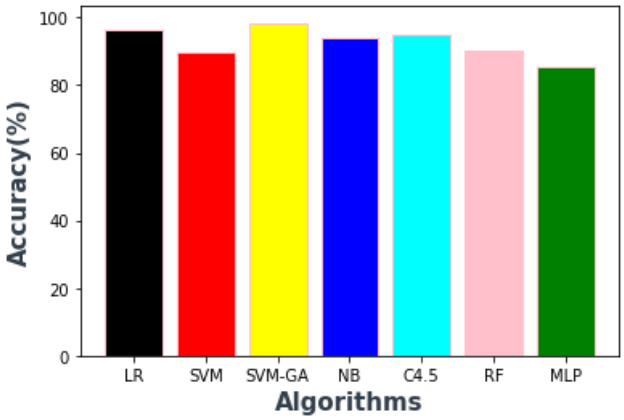

Fig. 3: Performance Comparisons with Accuracy

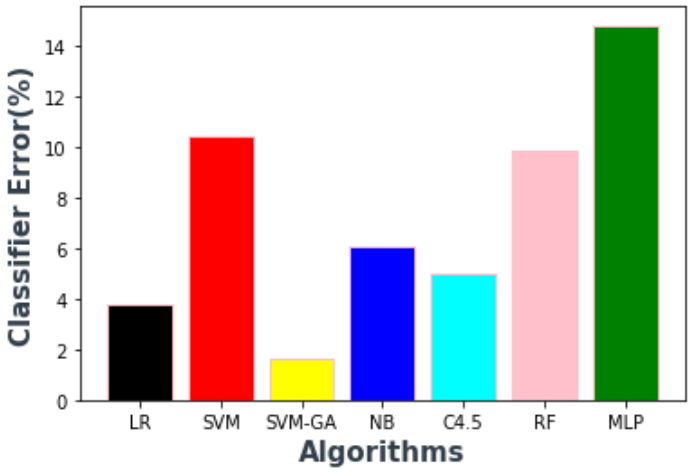

Fig. 4: Performance Comparisons with Classifier Error

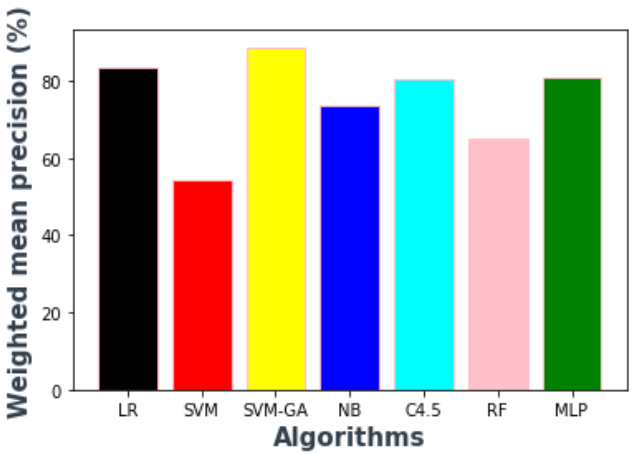

Fig. 5: Performance Comparisons with Precision

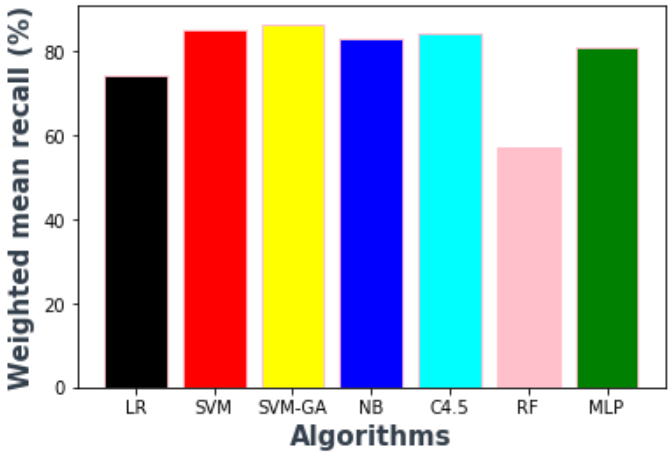

Fig. 6: Performance Comparisons with Recall

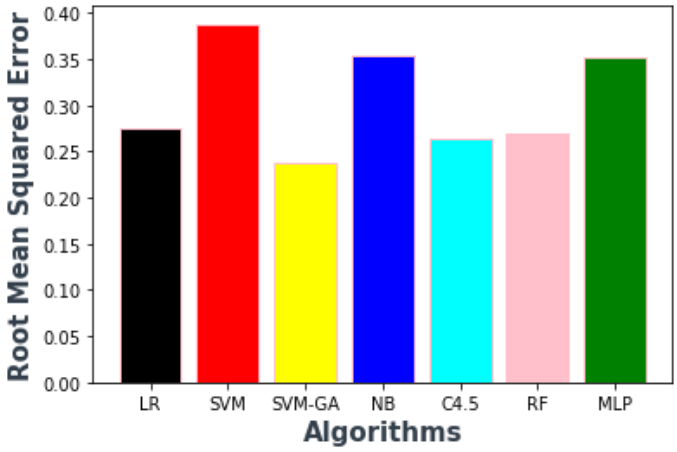

Fig. 7 Performance Comparisons with Root Mean Squared Error

8.0 CONCLUSION

Spam is a solemn issue that is not just exasperating to the end-users, but also financially damaging many industries. In this study, a deep review on state-of-the-art machine learning algorithms for spam filtering were done extensively and their field of application too. A variety of handful attempted

algorithms for spam filtration is narrated briefly. The challenges faced by the existing classifiers and the need for a hybrid is outlined. The process involved and the basic framework of spam message filters are provided. Experimentation is done with the well-known publicly available SMS spam dataset and the results were examined over existing machine learning classifiers. The results are recorded in the form of plots for different performance metrics such as recall value, precision, RMSE and accuracy. The proposed evolutionary based algorithm outperforms well compared to existing approached in terms of accuracy, precision, recall and RMSE.