Detecting Spam E-mails with Content and Weight-based Binomial Logistic Model

Richa Indu* and Sushil Chandra Dimri

Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India
E-mail: richaindu@outlook.com; dimri.sushil2@gmail.com
**Corresponding Author*

Received 05 May 2023; Accepted 08 December 2023;
Publication 03 February 2024

Abstract

Spam e-mails are continuously increasing and are a serious threats to a network and its users. Several efficient methods are available regarding this context, but still, it is evolving randomly. Considering this, the proposed approach addresses the problem of spam detection by combining traditional content-matching criteria with the modified version of the binomial logistic algorithm. The work generates seven categories for content-matching, which begins from three basic categories, namely: special words, adult content, and specific symbols and digits. The remaining four categories are derived from various possible combinations of these basic categories. The words selected for each category are carefully curated based on the human psychology of action and reaction. Then, a weight is assigned to each of the categories to signify their importance and a threshold criterion is deployed before implementing the binomial logistic algorithm, which not only increases the efficiency of the proposed algorithm but also reduces the rate of misclassification. The proposed model is tested on six separate datasets of Enron Spam Corpus, where 98.31% and 92.575% are the maximum and

Journal of Web Engineering, Vol. 22_7, 939–960.
doi: 10.13052/jwe1540-9589.2271
© 2023 River Publishers

minimum accuracies achieved, respectively, in spam e-mail classification. The AUC_ROC scores for the entire Spam Corpus range between 0.927 and 0.983. A comparison is also carried out between the proposed algorithm and the other methods of spam detection that have logistic regression. Finally, the suggested method can adequately handle a large sample size without compromising the efficacy, which is measured using accuracy, precision, recall, F-measure, and AUC_ROC score.

Keywords: Logistic regression, malicious advertisements, maximum likelihood estimation, spam e-mails.

1 Introduction

For more than a decade, the Internet has become a critical part of our daily lives. With increased dependency on the Internet, e-mails have become one of the popular means of communication. According to an estimation, 306.4 and 319.60 billion e-mails were sent every day during 2020 and 2021, respectively [1]. Such an increased usage of e-mails also paved the path for numerous profitable and commercial ways of advertising, even unsolicited ones, referred to as spamming. Statistical studies reveal that out of the total e-mails in a day around 85% are spam. Of these, 5.8% are fraudulent scams, 31.7% are related to adult content, 36% belong to the category of advertisements for various products and services, and remaining 26.5% are associated with financial matters [2].

Generally, spam or junk e-mails are those unwanted e-mails that originate from a random untrusted source from any location across countries and are broadcast to internet users via e-mail. Spam mail is intended to serve several purposes, some of these are for the promotional activities of products and related services (as advertisements), for identity and information theft (i.e., phishing attempts and malvertisements), and for launching malicious payloads (ransomware attacks) [3]. Moreover, spam sometimes leads to DDoS (Distributed Denial of Service) attacks by overwhelming the network bandwidth with traffic. The first known spam attack was in 1978 via an advertisement by Gary Thuerk [4]. However, spamming appeared as a serious issue during the 1990s with the commercialization of ARPANET as the Internet.

To address this emerging situation, several approaches were used. The most common and constantly evolving approach is spam filters. These spam filters use static and dynamic methods of spam detection [3]. The static methods are based on identifying illegitimate e-mail addresses from a predefined list. While the dynamic methods possess a self-learning behavior by scanning the occurrence of suspicious words in e-mails. Thus, preventing undesirable and annoying e-mails from reaching users' inboxes and restricting them to a separate folder called the junk or spam folder. However, with advancements in technology, detecting, classifying, and handling such e-mails has become difficult. Typically, to avoid spam filters, spammers continuously change their e-mail signatures. In this context, traditional static methods require speedy tuning to collect these new signatures, thus machine learning algorithms can be deployed on the server side for filtering.

Now, even network-providing companies deploy spam filters at their levels to reduce and rectify this problem in mobile networks. Several efficient methods are also available regarding spam detection and filtering, but it is evolving randomly. Therefore, a combination of machine learning approaches with traditional methods of spam detection can provide a more effective solution to such a problem of spam e-mails.

The main highlights of the proposed work are as follows:

- The proposed work uses content-matching criteria from the traditional approach of spam detection and creates seven different categories.
- The words selected for each category are carefully curated based on the human psychology of action and reaction.
- Weights are evaluated and assigned to each category signifying their importance in spam identification.
- Before implementing the binomial logistic algorithm for final classification, a threshold is computed to reduce the rate of misclassification.
- The efficacy of the proposed model is measured using accuracy, precision, recall, F-measure, and AUC_ROC (Area Under the Curve-Receiver Operating Characteristics) scores.

In short, instead of using any deep learning and evolutionary method, the present work modifies the binomial logistic regression method by incorporating weights and content filters for detecting spam e-mails without compromising with the efficiency of the model. The paper comprises six sections, where Section 2 provides a glimpse of related work and Section 3 discusses the proposed methodology. The information about the dataset used is included in Section 4, Section 5 represents and discusses results, followed by the conclusion in Section 6.

2 Related Work

Within the last decade several approaches have been applied to detect spam e-mails. One such was the content-based filtering deployed in 2015 by Rathod and Pattewar which scans only the body of the e-mail [5]. For detecting the probability of the e-mail being spam with a Bayesian classifier, they used three datasets containing a combination of 1000, 1500, and 2100 spam and ham e-mails. Their third set attained the maximum accuracy, precision, and recall scores of 96.46, 95, and 87%. Similarly, for detecting spam, Goh and Singh compared the performance of various machine-learning algorithms and meta-heuristic methods [6]. These machine-learning and meta-heuristic approaches were Artificial Neural Network (ANN), Support Vector Machine (SVM), Bayesian network classifier, Naive Bayes (NB), K-Nearest Neighbors (KNN) and Random Forest (RF), and AdaBoost, LogitBoost, Real AdaBoost, Bagging, Dagging, and rotation forest respectively. The two datasets used by them were WEBSPAM-UK2006 and 2007, where the maximum accuracy achieved was 93.7 and 85.2% on UK2006 and 2007 datasets respectively using AdaBoost with RF. Similarly, by combining LR with Logit, they achieved an efficiency of 93.1 and 84.7%, respectively, on the same datasets. However, out of the several other ensembles, only the RF with AdaBoost was identified as the best-performing pair for spam identification.

In the very next year, a hybrid model was proposed, which used Logistic Regression (LR) as a filter for Decision Tree (DT) [7]. The use of LR as a filter overcomes the limitation of DT for not handling noisy data. Furthermore, the introduction of a False Negative (FN) threshold between LR and DT increased the efficiency of their work for detecting spam e-mails from a sample size of 4601. Using this hybridized DT with LR and FN threshold, they attained 91.67% accuracy on the Spambase dataset. Again using the UCI Spambase dataset, another spam filtering model was designed [8]. It comprised LR with agglomerative clustering in two steps. Their proposed approach groups similar features according to computed weightage and emphasized analyzing spammers' patterns. They enhanced the classification accuracy from 93.03% (with all features) to 98.41% (with the feature selection approach). Again, using the same Spambase dataset, Bassiouni et al. tried to detect spam e-mails with the help of the infinite latent feature selection technique and classified those obtained features with 10 machine learning methods [9]. Out of those 10 supervised learning algorithms, the radial basis function kernel SVM has the lowest performance accuracy of 82.6%, whereas RF has the maximum efficiency of 95.45%. In addition, LR achieved an AUC_ROC score of 0.971 and an accuracy of 92.41%.

Likewise, Shah and Kumar used the genetic algorithm as the feature selector, not only to increase the efficacy of the LR classifier, but also to minimize the error rate [10]. With this bio-inspired algorithm, they increased the accuracy rate generated by the static methods from 86.73% to 88.931%. Another approach to the Spambase dataset used LR with the "select by weight" method to enhance the overall performance of identifying spam from the sample size of 4601 [11]. On 56 attributes, the gradient boost tree was used, which provided an accuracy of 95.13%. In an analytical study of finding spam from 2000 e-mails, RF outperforms LR and DT with an accuracy of 98% achieved in 0.19 seconds [12]. Nandhini and Marseline carried out a performance evaluation between LR, DT, NB, KNN, Random Tree (RT), and SVM for classifying spam e-mails from the Spambase dataset of the UCI Repository [13]. Their study showed that RT and KNN are the best spam classifiers with 99.94% accuracy, while only 93.13% accuracy is achieved with LR. Recently evolutionary algorithms have also been introduced as feature selectors for spam classification. Dedeturk and Akay used three different datasets; namely, Turkish Corpus, CSDMC2010, and Enron Spam Datasets [14]. Apart from highly efficient classification with LR, Gaussian, and multinomial NB, linear, and radial basis kernel SVM, they also addressed the issue of high dimensionality in such data. Thus, they applied Artificial Bee Colony Optimization (ABCO) for optimal feature selection. Each of their selected algorithms achieves accuracy above 90.93%. With LR, the classification accuracies on Turkish, Enron, and CSDMC2010 were 99.13, 98.20, and 98.31% respectively. And 98.31, 98.91, and 99% accuracies were achieved on the CSDMC2010, Enron dataset, and Turkish E-mail datasets respectively by ABCO-LR.

Kawale and Sait analyzed the whole network for evaluating spam [15]. For this, they acknowledged each network with weight and treated them as a node based on the review, functionality, and user behavior. Their work suggested a customized server filtering system as the best solution for e-mail filtering since this reduces cost, offers global access, and increases accuracy with less error. Recently, a renowned method for processing natural language problems, named the Term Frequency-Inverse Document Frequency (TF-IDF) approach, had been used [16]. This approach not only addressed the problem of the sparsity of data but also provided a mathematical significance to each word in the analyzed document. This technique, along with singular value decomposition (SVD), encoded each word in the Spambase dataset with a number and created a wordbook comprising various terms. Afterward, it calculates the frequency of those various terms and assigns a weight accordingly. This approach achieved an accuracy of 99.17% in spam e-mail identification. Despite this, the TF-IDF SVD method has a major limitation that it works only on the lexical level and is not capable enough to capture semantics. Similarly, Debnath and Kar applied deep learning methods, namely, Bidirectional Encoder Representations from Transformers (BERT), Bi-Long Short-Term Memory (BiLSTM) and LSTM on the Enron Spam Email Corpus and attained accuracy rates of 99.14, 98.34 and 97.15% respectively [17]. The Phish responder method with LSTM and multi-layer perceptron was suggested in [18], which on the Spambase dataset gained 99 and 94% accuracies respectively. Again on the same dataset, 97% accuracy was attained by Adam and RMS using prop-optimized LSTM [19]. Jilani and Sultana tried to classify emails as spam or ham based on Uniform Resource Locators (URLs) in the body of the message [20]. For this, they used and compared RF with other supervised machine-learning algorithms, and achieved 97% accuracy.

Recently, Sadia and colleagues carried out a comparative study of several Machine Learning (ML) approaches in identifying spam e-mails from the TweetR dataset [21]. In this analysis, 89% was the highest accuracy achieved by NB, while LR attained 85% accuracy only. A lightweight ML-based technique, which utilized word-frequency patterns, was suggested by Bouke and others for spam detection [22]. On classifying the Spambase data with RF and LR, the accuracies achieved were 97 and 92% respectively. Similarly, Das, Mandal, and Basak developed a three-parallel layered decision-based approach, which attained 98.4% accuracy in distinguishing spam from ham [23]. Zivkov and others deployed the evolutionary multi-verse optimizer swarm intelligence approach before classifying the CSDMC2010 dataset with LR to effectively separate spam emails [24]. Another evolutionary technique was suggested in [25], which used the atomic orbital search approach along with LR and gradient descent to improve the rate of detecting spam e-mails. On Enron (with 1000 samples) and CSDMC2010 (with 500 samples) datasets, they achieved an F-score of 78.33 and 96.30% respectively. Similarly, Al-Zoubi, Mora, and Faris utilized Harris Hawk optimization and weighted SVM for multi-lingual spam detection [26]. BERT, one-hot encoding, TF-IDF, and NGram methods were applied to obtain pre-trained word embeddings on Spanish, Arabic, and English language data corpuses. In contrast to other state-of-the-art approaches, their method achieved 88.163, 71.913, 89.565, and 84.27% accuracies on English, Spanish, Arabic, and multi-lingual corpuses, respectively. Sai and Swaminathan compared LR and KNN for classifying spam e-mails, where LR was 96% and KNN was 89% accurate [27]. Moutafis, Andreatos, and Stefaneas applied multiple ML algorithms on the Enron1 spam corpus [28]. Out of several ML approaches, SVM and LR gained the highest accuracies of 99.38 and 99.22% respectively.

However, the different needs of each individual regarding spam filters, and the high dimensionality of such data affects the efficacy of existing algorithms with low detection and satisfaction rates. Also, with deep learning methods, high-efficiency scores can be achieved, but such approaches consume more resources. Therefore, instead of using any deep learning method and even without compromising the efficiency of the model, our present work tried to resolve these issues by modifying the binomial logistic regression method by incorporating weights and content filters with it.

3 Proposed Methodology

3.1 Logistic Regression

Logistic Regression (LR) is a supervised machine learning algorithm that distinguishes two or more classes, depending on the type of problem. Unlike linear regression it is categorical, and the predicted output can either be ordinal, binomial, or multinomial [29, 30]. Mathematically, it derives probabilistic values confined between 0 and 1, from the equation of the line:

$$y = a_0 + a_1 x \tag{1}$$

where a_0 and a_1 are constants, and are generally referred to as the slope and intercept of the line. Now, y can yield any number depending on a_0, a_1, and x. But, according to our problem, it is required to restrict these continuous values of y between 0 and 1, such that it incorporates all the values between $-\infty$ and ∞, without decreasing the correlation among data points. Moreover, $y = 1$ represents that the considered e-mail is spam, whereas, $y = 0$ indicates the considered e-mail is non-spam. The function that can map y from (1) in the range $f\colon R \to (0, 1)$ is the logarithm of the ratio of the probability of success and that of failure, i.e., the logarithm of the odds ratio, given by:

$$\log\left(\frac{P}{1-P}\right) = a_0 + a_1 x \tag{2}$$

where P is the probability of getting a spam e-mail, and $1 - P$ is the probability of getting a non-spam e-mail. Further simplification of the above equation with the help of the exponential function provides

$$\frac{P}{1-P} = e^{a_0+a_1x} \tag{3}$$

On solving, it yields a sigmoidal function shown in (4),

$$P = \frac{e^{(a_0+a_1x)}}{1+e^{(a_0+a_1x)}} \tag{4}$$

or

$$P = \frac{1}{1+e^{-(a_0+a_1x)}} \tag{5}$$

In more generalized terms, (5) can also be written as $f(y) = \frac{1}{1+e^{-y}}$. Moreover, to evaluate the constants a_0 and a_1, the Maximum Likelihood Estimation (MLE) method is used [31, 32]. For a set of input features $X = x_1, x_2, \ldots, x_p$, y from (1) can be written as a linear function of these input features, i.e., $y = a_0 + a_1x_1 + \cdots + a_px_p$. In this case, we only have two parameters for value estimation, which are a_0 and a_1, and an outcome variable Y, whose value can either be p (i.e., 1 indicating spam e-mails) or $1 - p$ (i.e., 0 resembling non-spam emails).

Since Y follows a discrete Bernoulli's distribution, then its probability mass function can be expressed as:

$$\begin{aligned} P(Y=1|X) &= f(a_0+a_1x) \\ &= \frac{e^{(a_0+a_1x)}}{1+e^{(a_0+a_1x)}} \\ &= \left(\frac{e^{(a_0+a_1x)}}{1+e^{(a_0+a_1x)}}\right)^{Y}\left(1-\frac{e^{(a_0+a_1x)}}{1+e^{(a_0+a_1x)}}\right)^{1-Y} \\ &= \frac{e^{Y(a_0+a_1x)}}{1+e^{(a_0+a_1x)}} \end{aligned} \tag{6}$$

Now, for $y_i \epsilon Y$ and $x_i \epsilon X$, the generalized likelihood of (6) can be obtained as:

$$\mathcal{L}_n(Y_1, Y_2, \ldots Y_n, a_0, a_1) = \prod_{i=1}^{n} \frac{e^{y_i(a_0+a_1x_i)}}{1+e^{(a_0+a_1x_i)}} \tag{7}$$

Similarly, the log-likelihood of (7) can be computed as:

$$\log(\mathcal{L}_n(Y_1, Y_2, \dots Y_n, a_0, a_1)) = \sum_{i=1}^{n} y_i(a_0 + a_1 x_i) - \sum_{i=1}^{n} \log(1 + e^{(a_0 + a_1 x_i)}) \tag{8}$$

Setting the gradient descent of the log-likelihood function equal to the zero vector in (8) generates a pair of distinguished equations shown in (9) and (10), known as maximum likelihood estimators, whose solutions can be obtained by experimental analysis.

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{(a_0 + a_1 x_i)}}{1 + e^{(a_0 + a_1 x_i)}} = 0 \tag{9}$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{x_i e^{(a_0 + a_1 x_i)}}{1 + e^{(a_0 + a_1 x_i)}} = 0 \tag{10}$$

In this way, the *logit function* or logistic regression converts a straight line into an S-curve. However, this function cannot handle non-linear as well as complex problems and is also prone to overfitting.

3.2 Content-based Filtering

The words, phrases, alphanumerics, symbols, and numerics selected for filtering spam in the present work comprise urgent, unusual, disreputable, and overpromising money-saving offers. Such words, for instance, *100% cashback, free, 50% discounts, act now, call now, love, hate, lose weight, $,* etc., in general, induce a person to take swift actions in response to either the positive or negative impact generated on human psychology [33], which ends most of the time in a loss. Based on these psychological effects, we selected certain words for categorizing e-mails into spam or ham. Now, by introducing random words, digits, and altering vocabulary, spammers try to dodge content-based spam filters. Therefore, the present work not only scans the presence of any random sequence of digits but also various combinations of alphabets, words, and alphanumerics. We have considered different words, numerics, alphanumerics, and phrases as features and categorized them into seven categories derived from the three basic categories mentioned in Table 1. These primary categories are s_1 (special words), s_2 (adult content), s_3 (specific symbols and digits). Derived from these primary categories, other

Table 1 Categorization of words or phrases embedded in e-mails

S. no.	Categories	Words and Phrases
1	Special words	Award, lottery, quiz, bank account, digit, Save dollars, millions, euros, Americans, join, one hundred percent, Pennies, worth rupees, thousands, dollars, thank you, good luck, call, greetings, marketing, contact, puzzle, bankruptcy, winner, free, credit, reference, number, insurance, subscription, buy, direct, seen, buying, judgments, clearance, order status, shipped, shopper, additional, orders, status, income, be your own boss, compete for your business, double, extra, income, earn, per week, expect, based, employment, biz, Homebased, home, make, making, online, opportunity, degree, potential, earnings, university, diplomas, work, affordable, bargain, deal, beneficiary, best price, bonus, big bucks, cash, hidden, assets, cents, cheap, check, claims, collect, loans, compare, rates, investment, bureaus, discount, easy terms, fast, save, serious, charges, incredible, lowest, no fees, back, mortgage, day, profit, pure, quote, refinance, refund, big, unsecured, debt, pay, more, accept, credit, debit, cards, accepted, check, cheque, offers, money, explode, business, full, investment, decision, hidden, requires, compliance, stock, alert, pick, disclaimer, calling, creditors, collect, child support, consolidate, eliminate, bad finance, get out, paid, lower rate, monthly, payment, lowest, pre-approved, social, security, number, chance, reverse, ad, appliances, acceptance, avoid, auto e-mail removal, dormant, freedom, leave, lose, lifetime, maintained, medium, miracle, passwords, problem, remove, sample, solution, success, stop, bulk, click below, here, direct, marketing, harvest, form, increase sales, traffic, market, mass, member, trial, multi-level, not spam, one time, opt, open, undisclosed recipient, unsubscribe, visit, website, web traffic, believe, billing address, shipping address, orders, gift, give away, confidential, believe, performance, junk, search engine, follow, certificate, turned down, important, information, regarding, laws, distance, mail, message, name, brand, restrictions, experience, catch, disappointment, gimmick, middleman, attached, no obligation, inventory, priority, prize, weekly mail, shopping spree, stuff, terms and conditions, vacation, warranty, wins, winner, off shore, unlimited, compare, copy, get, print, signature, fax, today, sign up for free, DVD, grant, hosting, installation, install now, getaway, honor, weekend, won, selected, time, cell phone, instant, real, risk, amazing, membership, consultation, natural, fantastic deal, certified, congratulations, promise, drastically reduced, apply now, can't live without, don't delete, preview, hesitate, illegal, legal, once in a lifetime, one time, please read, luxury car, domain extension, casino, celebrity, great offer, special promotion, limited supply, urgent action needed, CD, pager, cable, converter, addresses, laser printer, Rolex, stainless teel, call now.

(*Continued*)

Table 1 Continued

S. no.	Categories	Words and Phrases
2	Adult content	Sleep, satisfaction, Nigerian, age, images, sex, content, Viagra, relation, medicines, apparatus, dating, capsules, dig up, dirt, dirty, meet friends, single, score with babes, teen, wife, hot, nude, dear friend, hello, cure baldness, adult, diagnostic, fast, love, hate, human growth hormone, life, lose weight, medical, pharmacy, remove wrinkles, reverse aging, stop snoring, Valium, Vicodin, weight loss, private, Xanax
3	Specific symbols and digits	$, €, ₹, 1,00,000, @, #, *, %, $XXX, XXX, #1, 100%, 4U, 50%, 25%, 20%, !

categories are obtained by combining s_1 and s_2 in the category s_4, s_1 and s_3 in the category s_5, and s_2 and s_3 in the category s_6. Finally, s_1, s_2, and s_3 are amalgamated into the category s_7.

3.3 The Proposed Algorithm

The foremost task is to compare the subject and content of each e-mail, word by word, with the categories of words mentioned in Section 3.2. Thereafter, the occurrence of words specific to the *i*th category, i.e., x_i is evaluated. Depending on the data, weights are assigned to each category according to (11), signifying their importance in spam e-mail identification.

$$w_i = \left(10 \times \frac{1}{\rho_i}\right) \tag{11}$$

where w_i stands for the weight assigned to the *i*th category, $\rho_i = \frac{s_i}{T_S}$, with s_i the number of spam e-mails due to the virtue of category i, and T_S is the total number of spam e-mails.

Then, for each e-mail, (1) can be transformed as $z = a_0 + a_1 X$, where X is computed as

$$X = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6 + w_7x_7 \tag{12}$$

$$f(z) = \frac{1}{1 + e^{-z}} \tag{13}$$

Now, the value z can vary from $-\infty$ to $+\infty$. Thus, to restrict it within the limits of 0 and 1, the logit function defined in (13) is used. Based on the experimental data, the constants a_0 and a_1 are evaluated using Equations (14) and (15) as derived from Equations (9) and (10), where $\overline{Z}$ and $\overline{X}$ are the means of Z and X respectively, such that $z_i \in (0, 1)$.

$$a_1 = \sum_{i=1}^{n} \frac{(x_i - \overline{X})(z_i - \overline{Z})}{(x_i - \overline{X})^2} \tag{14}$$

$$a_0 = \overline{Z} - a_1\overline{X} \tag{15}$$

Then, from the series of experimentally obtained values of a_0 and a_1, the best pair is selected. This optimal pair is the one, which maximizes the performance of the proposed algorithm on that particular dataset.

Finally, the evaluated $f(z)$ is compared with the threshold. Now, if $f(z) < 0.5$, then the predicted output is transformed to 0 representing ham (non-spam). Otherwise, if $f(z) \geq 0.5$, then the predicted output is set to 1 denoting spam e-mail.

The summarized version of the proposed algorithm is shown below:

Proposed Algorithm for Spam Detection:

Input: E-mails as text files.

Output: [0,1]; where 0 stands for non-spam, and 1 for spam e-mail.

Step 1: Read the e-mails one by one.

Step 2: Scan for the occurrence of each word phrase, and count the incidence of each word from each of the seven categories mentioned in Section 3.2, i.e., (x_i) for $i = 1$ to 7.

Step 3: Calculate weight for each category using (11).

Step 4: Compute X using (12).

Step 5: Substitute value of X in the transformed equation of linear regression, i.e., $z = a_0 + a_1X$.

Step 6: Calyculate and select the optimal values of a_0 and a_1, using (14) and (15).

Step 7: Evaluate the logit function mentioned in (13).

Step 8: Compute the final output by comparing the outcome of step 7 with the threshold limit as:

If $f(z) \geq 0.5$, then output = 1 (Spam),

Otherwise, if $f(z) < 0.5$, then output = 0 (Non-spam).

4 Dataset Used

The Enron dataset is used for measuring the efficiency of the proposed algorithm [34, 35]. There are six separate directories of datasets in the Enron spam corpus, numbered from 1 to 6. Each preprocessed subdirectory contains messages in the text file format, where each filename begins with the order of arrival, which contains the e-mails from the senior management of the Enron Corporation. The total sample size of the Enron spam corpus is 30,703, where 17,170 and 14,033 e-mails are labeled as spam and non-spam respectively. Moreover, the first dataset comprises 1500 spam and 3672 non-spam e-mails, thus a total of 5172 e-mails. Similarly, the second dataset consists of 5857 e-mails, where 1496 are spam and 4361 are ham. Likewise, the rest of the datasets include 3000 (1500 spam and 1500 non-spam e-mails), 6000 (out of which 4500 are spam and the rest are non-spam), 5175 (3675 spam and 1500 ham), and 5999 (4499 e-mails are spam and rest are non-spam) e-mails.

5 Results

Being a content and weight-based algorithm, the proposed model effectively uses the words mentioned in Section 3.2, and weights associated with each category to identify the spam e-mails with the help of the binomial logistic function. For testing the efficiency of the present work, a large sample size of 30,703 is used in six parts, with 5172, 5857, 3000, 6000, 5175, and 5999 e-mails in datasets 1 to 6 respectively. For the largest sample of 6000 e-mails, the proposed work seeks 0.053 seconds for processing one e-mail despite its high computational complexity. Likewise, the maximum and minimum accuracy of the proposed spam identification method is 98.31% and 92.575% for datasets 5 and 1 respectively. Therefore, based on the aforementioned outcomes these datasets can be sorted in the descending order of performance, i.e., from maximum to minimum efficiency rates, like $D5 > D3 > D6 > D4 > D2 > D1$, where D represents a dataset.

The results so obtained are promising on the total sample size, where the overall average of accuracy, precision, and recall are 95.2875, 96.23, and 92.5783% respectively. Similarly, the harmonic mean of the precision and recall score, i.e., F-measure is 94.2583% on average, indicating good performance on an imbalanced classification problem. These various performance measures are graphically depicted in Figure 1 for individual datasets. In comparison with the earlier techniques, the results revealed that the proposed algorithm works efficiently on a large sample size without using any feature selection or evolutionary algorithm. The inclination of both recall and precision toward 1 signifies that the model does not miss any true positives.

As evident from Figure 1 also, the highest and the lowest F-measures are 98.81% (for dataset 5) and 87.91% (for dataset 1) respectively. Similarly, the

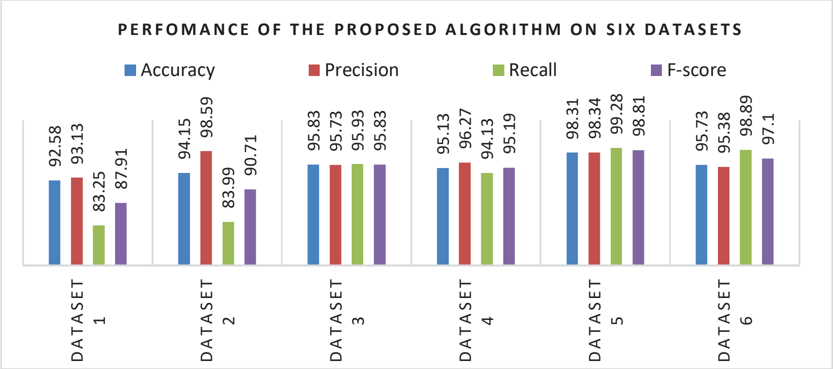

Figure 1 Performance of the proposed algorithm on six datasets indicating accuracy, precision, recall, and f-score of the proposed model.

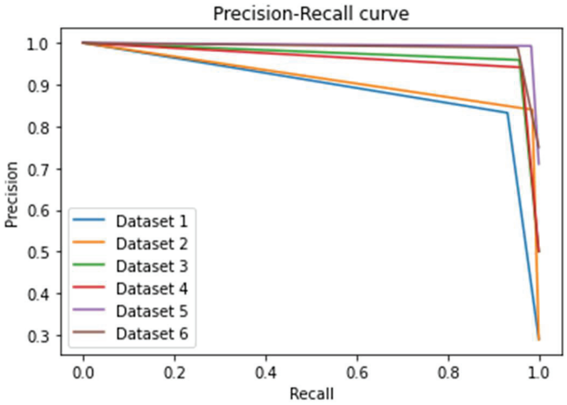

Figure 2 Precision–recall curves of each dataset in the Enron spam e-mail corpus signifying the authenticity of the present work in detecting spam with the least misclassification rate.

precision score of the proposed approach is always above 90%. Meanwhile, the recall score ranges from 83.25% to 99.28%. Precision focuses on type-I error, which is equivalent to rejecting a true null hypothesis. In simpler words, any incorrect classification of spam as non-spam e-mail contributes to type-I error. Unlike precision, recall (or sensitivity) concentrates on type-II errors, which corresponds to accepting a false null hypothesis. Simply, a type-II error is the misclassification of non-spam e-mail as spam e-mail. This model also reduces the rate of misclassification, where the large area under the precision–recall curve signifies both high precision as well as recall. It is also evident from the precision–recall curve shown in Figure 2.

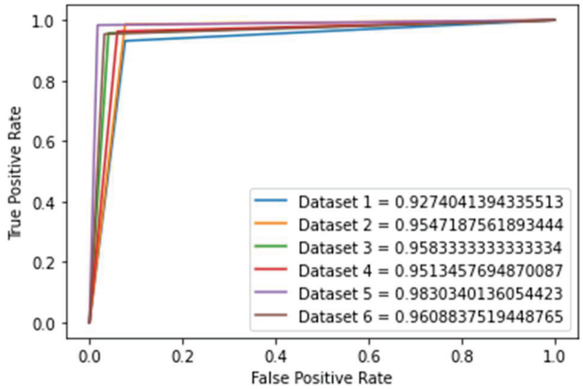

Figure 3 AUC_ROC curves and scores of each dataset in the Enron Spam Corpus indicating the efficiency of the proposed work in identifying the two classes, i.e., spam and non-spam.

Likewise, the trade-off between the true and false positive rate is represented by the receiver operating characteristic (ROC) curves graphically, and the area under this curve (AUC) provides a significant illustration of the capability of the proposed approach to distinguish between the two classes, i.e., 1 (for spam e-mails) and 0 (for non-spam e-mails). The probability curve for each of the six datasets is depicted in Figure 3 along with the obtained AUC scores. The AUC_ROC scores in decreasing order are 0.983 (dataset 5), 0.96 (dataset 6), 0.958 (dataset 3), 0.954 (dataset 2), 0.951 (dataset 4), and 0.927 (dataset 1).

Figure 4 comprises a comparative analysis between the present approach, and other methods using logistic regression for spam detection during 2015–2023. These methods used Web Spam UK2006, UK2007, Spambase, SpamURLs, CSDMC2010, TweetR, and Enron Spam Corpus datasets. In contrast with other algorithms [6–11, 13] the accuracy ranges from 84.7 to 95.13%. While [14] (LR without ABCO) and the proposed approach has a comparative efficiency of 98.20 and 98.31% respectively. However, when LR is used with ABCO on the same dataset their accuracy rises to 98.91%. Meanwhile, the methods that performed a bit better in comparison to the suggested approach used deep learning algorithms. Thus, instead of using deep learning methods like BERT and LSTM, and evolutionary algorithms such as ABCO, the proposed effort used merely a machine learning algorithm, namely, binomial logistics regression, with weights and content filters to detect spam emails. Besides this, the suggested approach can efficiently and precisely handle large datasets, and also reduce the misclassification instances

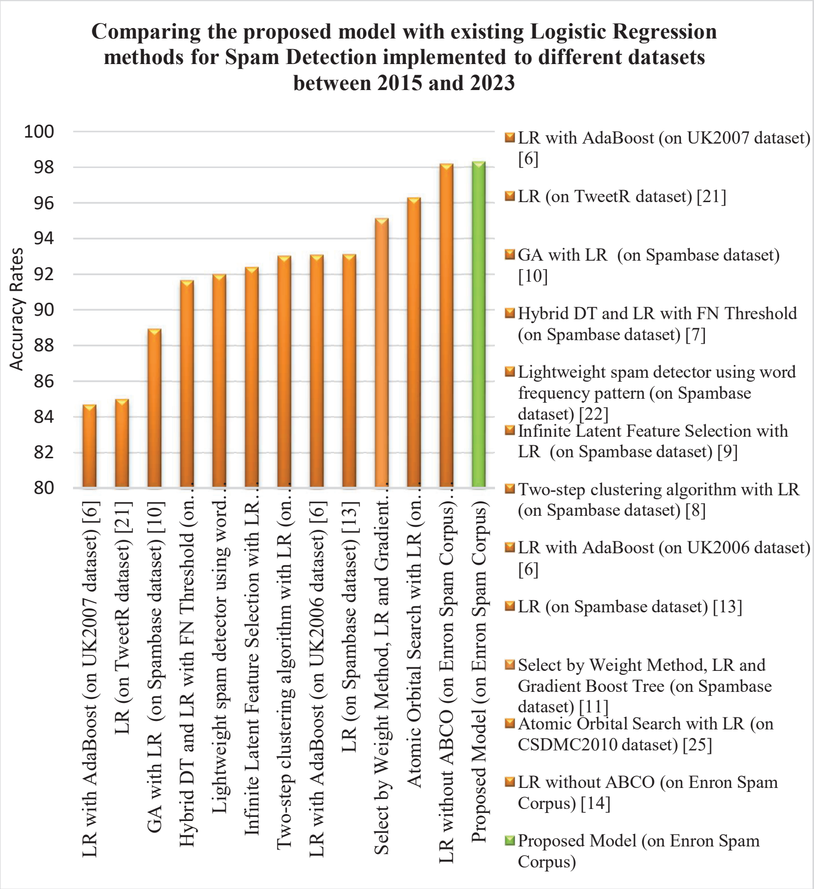

Figure 4 A comparison between the proposed spam detection model and the existing LR methods of spam filtering, applied to different datasets between 2015 and 2023.

by applying a threshold. Further, for computing the smallest sample size of 1500 e-mails, it takes only 0.0003 seconds.

6 Conclusion

Spam e-mails are serious threats not only to the client systems but also to the network infrastructure. An effective solution to this rising problem is suggested as a content and weight-based algorithm. Inspired by the traditional approach, this algorithm designs seven content-specific categories from three basic categories, namely special words, adult content, and specific symbols and digits, where each word is carefully curated according to the category. Then, to improve the overall efficiency of detection, assessed weights are assigned according to the significance of the category in spam detection. The final classification outcome is generated using the binomial logistic algorithm after cross-verifying each prediction with the threshold value to reduce the misclassification rate. The present work attains the highest accuracy, precision, recall, and F-measure of 98.31, 98.34, 99.28, and 98.81%, respectively, on dataset 5 of the Enron Spam Corpus. Similarly, the outcomes on the remaining datasets also acknowledge the effectiveness of the proposed approach, where the minimum accuracy, precision, recall, and F-measure are 92.58, 93.13, 87.91, and 87.91%, respectively for dataset 1. Further for all datasets, the AUC_ROC scores of the suggested method range between 0.927 and 0.983. Therefore, the proposed method can efficiently handle a large sample size and takes only 0.0003 seconds for computing the smallest sample size of 1500 e-mails. Additionally, apart from the body of the e-mail, the presented algorithm also considers the subject of the e-mail for spam detection. In the future, we will further try to improve the classification of the algorithm with other methods and also include the attached documents for the process of spam detection.

References

[1] J. Johnson, 'Number of sent and received e-mails per day worldwide from 2017 to 2025', Statista Research Service, 2021. https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/.

[2] N. Cveticanin, 'What's on the Other Side of Your Inbox – 20 SPAM Statistics for 2022', DataProt, 2022. https://dataprot.net/statistics/spam-statistics/.

[3] R. Indu, A. Sharma, 'Ransomware: A New Era of Digital Terrorism', Computer Reviews Journal, vol. 1, no. 2, pp. 168–226, 2018.

[4] G. Vijayasekaran, S. Rosi, 'Spam And E-Mail Detection in Big Data Platform Using Naive Bayesian Classifier', Int. J. Comput. Sci. Mob. Computing, vol. 7, no. 4, pp. 53–58, 2018.

[5] S.B. Rathod, T.M. Pattewar, 'Content based spam detection in email using Bayesian classifier', Proc. In ICCSP, pp. 1257–1261, Melmaruvathur, India, 2015. https://doi.org/10.1109/ICCSP.2015.7322709.

[6] K.L. Goh, A.K. Singh, 'Comprehensive Literature Review on Machine Learning Structures for Web Spam Classification', Procedia Comput. Sci., vol. 70, pp. 434–441, 2015. https://doi.org/10.1016/j.procs.2015.10.069.

[7] A. Wijaya, A. Bisri, 'Hybrid decision tree and logistic regression classifier for e-mail spam detection', Proc. In ICITEE, pp. 1–4, Yogyakarta, Indonesia, 2016. https://doi.org/10.1109/iciteed.2016.7863267.

[8] A.H. Osman, H.M. Aljahdali, 'Feature Weight Optimization Mechanism for Email Spam Detection based on Two-Step Clustering Algorithm and Logistic Regression Method', Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 10, pp. 420–429, 2017.

[9] M. Bassiouni, M. Ali, E.A. El-Dahshan, 'Ham and Spam E-Mails Classification Using Machine Learning Techniques', J. Appl. Secur. Res., vol. 13, no. 3, pp. 315–31, 2018. https://doi.org/10.1080/19361610.2018.1463136.

[10] N.F. Shah, P.A. Kumar, 'Comparative Analysis of Various Spam Classifications', In P. Sa, M. Sahoo, M. Murugappan, Y. Wu, B. Majhi (eds.) Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Advances in Intelligent Systems and Computing. vol. 719, Springer, Singapore, 2018. https://doi.org/10.1007/978-981-10-3376-6_29.

[11] A. Anggraina, R. Primartha, A. Wijaya, 'The Combination of Logistic Regression and Gradient Boost Tree for Email Spam Detection', In IOP Conf. Series: Journal of Physics: Conf. Series, vol. 1196, pp. 012013, 2019. https://doi.org/10.1088/1742-6596/1196/1/012013.

[12] B. Santoso, 'An Analysis of Spam E-mail Detection Performance Assessment Using Machine Learning', Jurnal Online Informatika, vol. 4, no. 1, pp. 53–56, 2019. https://doi.org/10.15575/join.v4i1.298.

[13] S. Nandhini, K.S. Marseline, 'Performance Evaluation of Machine Learning Algorithms for E-mail Spam Detection', Proc. In ic-ETITE, pp. 1–4, Vellore, India, 2020. https://doi.org/10.1109/ic-etite47903.2020.312.

[14] B.K. Dedeturk, B. Akay, 'Spam filtering using a logistic regression model trained by an artificial bee colony algorithm', Appl. Soft Comput., vol. 91, 2020. https://doi.org/10.1016/j.asoc.2020.106229.

[15] N.J. Kawale, S.Y. Sait, 'A Review on Various Techniques for Spam Detection', Proc. In ICAIS, pp. 1771–1775, Coimbatore, India, 2021. https://doi.org/10.1109/icais50930.2021.9395979.

[16] W. Park, N.M.F. Qureshi, D.R. Shin, ‘Pseudo nlp joint spam classification technique for big data cluster’, Computers, Materials & Continua, vol. 71, no. 1, pp. 517–535, 2022.

[17] K. Debnath, N. Kar, ‘Email Spam Detection using Deep Learning Approach’, COM-IT-CON Int. Conf. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, pp. 37–41, Faridabad, India, 2022. https://doi.org/10.1109/COM-IT-CON54601.2022.9850588.

[18] M. Dewis, T. Viana, ‘Phish Responder: A Hybrid Machine Learning Approach to Detect Phishing and Spam Emails’, Appl. Syst. Innov., vol. 5, no. 4, 2022. https://doi.org/10.3390/asi5040073.

[19] E. John-Africa, V.T. Emmah, ‘Performance Evaluation of LSTM and RNN Models in the Detection of Email Spam Messages’, EJCSIT, vol. 2, no. 6, pp. 24–29, 2022. http://dx.doi.org/10.24018/ejcompute.2022.2.6.80.

[20] A.K. Jilani, J. Sultana, ‘A Random Forest Based Approach to Classify Spam URLs Data’, ICETSIS Int. Conf. Emerging Technologies for Sustainability and Intelligent Systems, pp. 268–272, Manama, Bahrain, 2022. https://doi.org/10.1109/ICETSIS55481.2022.9888849.

[21] A. Sadia, F. Bashir, R. Q. Khan, and A. Khalid, ‘Comparison of Machine Learning Algorithms for Spam Detection,’ Journal of Advances in Information Technology, vol. 14, no. 2, pp. 178–184, 2023. http://dx.doi.org/10.12720/jait.14.2.178-184.

[22] M. A. Bouke, A. Abdullah, M. T. Abdullah, S. A. Zaid, H. El Atigh, and S. H. ALshatebi, ‘A Lightweight Machine Learning-Based Email Spam Detection Model Using Word Frequency Pattern,’ Journal of Information Technology and Computing, vol. 4, no. 1, pp. 15–28, 2023. http://dx.doi.org/10.48185/jitc.v4i1.653.

[23] S. Das, S. Mandal, and R. Basak, ‘Spam email detection using a novel multilayer classification-based decision technique,’ International Journal of Computers and Applications, vol. 45, no. 9, pp. 587–599, 2023. http://dx.doi.org/10.1080/1206212X.2023.2258328.

[24] M. Zivkovic, A. Petrovic, N. Bacanin, M. Djuric, A. Vesic, I. Strumberger, and M.Marjanovic, ‘Training Logistic Regression Model by Hybridized Multi-verse Optimizer for Spam Email Classification,’ Proceedings of International Conference on Data Science and Applications, pp. 507–520, 2023. http://dx.doi.org/10.1007/978-981-19-6634-7_35.

[25] G. Manita, A. Chhabra, and O. Korbaa, ‘Efficient e-mail spam filtering approach combining Logistic Regression model and Orthogonal Atomic

Orbital Search algorithm,' Applied Soft Computing, vol. 144, 2023. http://dx.doi.org/10.1016/j.asoc.2023.110478.

[26] A. M. Al-Zoubi, A. M. Mora and H. Faris, 'A Multilingual Spam Reviews Detection Based on Pre-Trained Word Embedding and Weighted Swarm Support Vector Machines,' in IEEE Access, vol. 11, pp. 72250–72271, 2023. https://doi.org/10.1109/ACCESS.2023.3293641.

[27] B. N. Sai, B. Swaminathan, 'Using the K-Nearest Neighbors Algorithm and Logistic Regression to Improve Accuracy, a Novel Machine Learning Approach for Detecting SMS Spam Message,' Journal of Survey in Fisheries Sciences, vol. 10, no. 1S, pp. 2831–2842, 2023. https://doi.org/10.17762/sfs.v10i1S.516.

[28] I. Moutafis, A. Andreatos, and P. Stefaneas, "Spam Email Detection Using Machine Learning Techniques," European Conference on Cyber Warfare and Security, vol. 22, no. 1, pp. 303–310, 2023. http://dx.doi.org/10.34190/eccws.22.1.1208.

[29] M. Maalouf, 'Logistic regression in data analysis: an overview', Int. J. Data Anal. Tech. Strateg., vol. 3, no. 3, pp. 281–299, 2011. https://doi.org/10.1504/IJDATS.2011.041335.

[30] IBM, 'What is logistic regression? – Learn how logistic regression can help make predictions to enhance decision-making', 2022. https://www.ibm.com/in-en/topics/logistic-regression.

[31] R. Febrianti, Y. Widyaningsih, S. Soemartojo, 'The parameter estimation of logistic regression with maximum likelihood method and score function modification', Proc. In BASIC, vol. 1725, pp. 012014, Depok, Indonesia, 2018. https://doi.org/10.1088/1742-6596/1725/1/012014.

[32] N. Agrawal, 'Decoding Logistic Regression Using MLE. Data Science Blogathon', Analytics Vidhya, 2022. https://www.analyticsvidhya.com/blog/2022/02/decoding-logistic-regression-using-mle/.

[33] J. Billieux, A. Heeren, L. Rochat, P. Maurage, S. Bayard, R. Bet, et al, 'Positive and negative urgency as a single coherent construct: Evidence from a large-scale network analysis in clinical and non-clinical samples', Journal of personality, vol. 89, no. 6, pp. 1252–1262, 2021. https://doi.org/10.1111/jopy.12655.

[34] Spam Enron Corpus Dataset. http://nlp.cs.aueb.gr/software_and_datasets/Enron-Spam/index.html.

[35] V. Metsis, I. Androutsopoulos, G. Paliouras, 'Spam Filtering with Naive Bayes – Which Naive Bayes?', Proc. In CEAS, pp. 28–69, Mountain View, California, USA, 2006.

Biographies

Richa Indu is currently pursuing a Ph.D. in Computer Science and Engineering from Graphic Era Deemed to be University, Dehradun. She accomplished M.Tech (Hons) in Computer Science and Engineering from Uttarakhand Technical University, Dehradun and M.Sc. (Gold medalist) in Information Technology from Hemavati Nandan Bahuguna University, Srinagar (Garhwal), India. She has published six papers in conferences and journals. Her research interest includes machine learning, programming languages, data sciences and designing algorithms.

Sushil Chandra Dimri is currently serving Graphic Era Deemed to be University as a professor in the CSE Department. He received an M.Tech. from IIT Dhanbad and a Ph.D. in Computer Science from Kumaon University, Nainital, Uttarakhand, India. He has 22 years of experience in teaching of UG and PG level degree courses. He is the author of many books and has published more than 60 papers in national/international conferences and journals. His areas of interest are algorithm design, resource optimization, machine learning and computer graphics.