A Review On Machine Learning Algorithms In Enhancing Email Security

[*1]Mrs Swapna Vanguru, [2]A. Kethsy Prabavathy

[*1]Research scholar, Department of Computer Science, Karunya University, Coimbatore, India.
[2]Assistant Professor, Department of Computer Sciences Technology, Karunya University, Coimbatore, India.
[*1]Corresponding Author: **[1]Mrs V Swapna**
Assistant Professor, Department of Computer Science & Engineering, Keshav Memorial Engineering College, Hyderabad.

How to cite this article: Swapna Vanguru, A. Kethsy Prabavathy (2024) A Review On Machine Learning Algorithms In Enhancing Email Security. *Library Progress International,* 44(3), 15898-15905

Abstract:
Electronic mail is one of the rapid and technological ways to send important messages from one location to another place worldwide. The development of the usage of the online mail has evolved to achieve the piece of information in the pillar box, where the beneficiary gets a huge number of information, most of them create significant and diversified challenges including the challenges related to the identity of the receivers, the abruption of crucial information and destruction of Internet. These communicational bridges are so harmful that the users cannot ignore that during their usage of various types of forms including the advertisements and other electronic messages. The Trash among the messages is the process of scrutinizing those harmful messages that are harmful for noticing them and creating the communication among viewers. This particular study explains the process in improving the fraudulent activities in investigating and by providing the proposed single objective to develop the challenges in creating algorithm that uses the "Machine Learning classification algorithms" in creating the regressive in making the classifications to build the "optimal model" for accurately categorizing the electronic messages. Cleaning the messages and selecting them for creating new features are the initial activities in modeling the processes to diminish the dimensions of "sparse text features" obtained from trash and "ham communications." The selection of various features is used to create the best dimensions and the third stage is to discover the brouches using a "Logistic Regression classifier" with accuracy level of that.

Keywords: Trash, Database, Choosing of Features, Regression of Logistics.

I . Introduction

Junk email has been identifying the most significant cyber-attacks within the recent few years. Spam is the communication process of unpopular and undemanded messages delivered by a sender to define the relationship with the receivers in an uninvited way [1]. There are various types of trash within the online correspondence including the "e-mail, SMS, social media, and online commerce platforms." The users have to spend a huge time as they have to recognize those fraudulent messages and remove those uninvited messages. Those unrelated messages devour the free space of chat boxes and create issues in maintaining the important messages in mail-id [2]. Besides this, the mailshot can be easily sent through the Internet of personal phones [3]. The increasing rate of the illegal people who are associated with these types of works and the aftermath of those works within the digital marketing system, hamper the ways of security concerns of the users. Due to the abruption in the increasing rate of this fraudulent works, the spam issues in "social networking sites" gain the attention of investigators and the users [4]. Thus, the process of percolating those unrelated messages is providing the important vision for relocating the routs of communication.

The investigation of those illegal activities can be located both mechanically and with persons. The filtration of trash electronic messages that involves the process of time-consuming and the engaging persons within the process are the most important aspect of this system. Moreover, the interconnection of those users and the participants have to experience various challenges in creating the communication system that create various challenges and among them, the "links to phishing" the "websites or malware-hosting" and those servers of internet of phones. Indicating the issues of electronic mail, various investigators and researchers have participated within this process in developing the automation process of "filtration" of those unwanted messages. Among various systems and tools present within the digital era, the "Machine Learning Algorithms" are one of the most important tools in detecting the trashes in electronic mail systems. The elementary factor of the process is to create a “word list” and providing weightage to the words. Moreover, the illegal threads incorporate with the mostly used “common statements” to eliminate the similarities within the “Neutral Networks” [5] that provide assistance to the “support vector machines (SVMs)” [6], “Naive Bayes (NB)” [7] [8], and “Random Forest (RF)” that the most significant techniques of “Machine learning” and the “filtration process of spam” [9].

As per the studies of Kaur [10], the “Assembling learning approaches” include the “bagging” and “Random Forest outperform classifiers” are the stages of traditional single. There are simple differences between the “single algorithms” and “ensemble processes” are the integration of predictions of various approaches of the “Machine Learning” to improve the precision and the accuracy level of those. Past researchers had shown that various types of “traditional classifiers” including the “decision trees” were used significantly within the approaches of ensemble. The limitations are within the less number of researches related to the “Neutral Networks (NNs)” within the learning process of ensemble.

This study has the main target in examining the models of “machine learning” on the basis of the “traditional algorithms” and ensembles with the use of a feature of representing the “spam filtering” after using the “open-source spam” datasets including “Enron and Spam Assassin.” Existing System

Several types of spamming emails are found that includes filtering the existing methods. The paper discusses the artistic state and categorises the email spams. The different types of Email Spam filtering techniques are categorised based on their wide usage. One of them is spam filtering based on contents and another one is spam filtering technique based on the case specific analysis.

Content dependent spam extraction method: Spam extraction depending on the contents generally includes automatic filtering rule creation for conducting email classification. The process uses the Machine Learning algorithms including the classification based on “Naive Bayesian method” along with the “K neural Network” algorithm. Moreover, the Supporting Vector based Machine is generated that mainly uses words based on the phrases of the content emails. Incoming spam email sets are isolated using this method for rule generating [28].

● Case dependent Extraction of Spam email: The categorised Case dependent extraction of spam emails is the most popular method of email spam extraction. The initial step includes extracting both spam and non-spam set of emails from the inbox of the users through the collection model. The step is followed by pre-processing technique using the client interface method. Along with these, feature extraction based proper selection of emails and grouping them by evaluating the data is typically important step that helps in categorising those data in two volatile sets of non-spam and spam emails. Finally, the datasets training is conducted using machine learning algorithm for testing them depending on the decision-making algorithm.

II. Related Work

Depending on the increasing rate of spam contents in the Internet, new combatting strategies for stopping the spamming processes are developed. The most primitive effort regarding detection of spam in automated way using the Machine Learning algorithms is the process of classification based on Bayesian network (Friedman, Geiger, & Goldszmidt, 1997) and (Sahami, Dumais, Heckerman, & Horvitz, 1998). Many publications including (Dalvi et al., 2004), (Chinavle, Kolari, Oates, & Finin, 2009), (Biggio, Fumera, & Roli, 2009) are present that help in development of adversarial strategies. Numerous sets of studies have been found regarding the evolving process of email spam creation including The current study has typical similarities with the articles written by (Guerra, et al., 2010) and (D. Wang et al., 2013) and (Dalvi et al., 2004), (D. Wang et al., 2013) and (Guerra, et al., 2010).

(Dalvi et al., 2004) have opined in their study that a significant amount of problems are present that are faced by the adversaries regarding adoption of classifiers in producing the false negative options. The researchers focused on creating an usage of the “Naive Bayesian Classifier” that focuses on extending the functionality of adversary tackling. On the other hand, (D. Wang et al., 2013) have focused on elaborating the dynamic email spamming nature for statistical way exploring. The main aim of the study included the process of email spam type evolution over the era. The most significant modelling algorithm based on the case that is used by the researchers is The Topic Modelling Algorithm. The Topic drift is conducted by the Topics of Email that focuses on performing proper analysis of the Adversary networks. Another important study is conducted by (Guerra, et al., 2010) that focused on analysing the spam filter importance upon spam controlling and evolving the spams. Moreover, they have explored the significance of spam datasets in finding the adversary credibility regarding spam extraction conduction. The dynamic nature of spam emails have been evaluated by the researchers in rule set analysis based on different Spam Assassin rule sets. The Spam Assassin is an open access Spam filtration platform that provides dataset testing a higher opportunity.

The past processes of spam extraction based on filtration method included different classifications like email spamming, web spamming and sms spamming. Phishing is considered as one of the most common process of spamming based Cyber attack, and the process includes stealing of personal and sensitive sets of information including the login credentials, passwords and different types of sensitive information in Credit Card numbers and other types of important credentials. The research study byhas opined in their study that phishing attacks significantly get recognised based on the level of authority. Machine learning is widely used for creating learning analysis of the user generated features that includes IP based URLs. The domain names are interconnected by the means of non modal websites that significantly consists of non modal Websites. HTML designing of emails varies based on domain names, along with number of dots and spam filter output. The proposed method by the researchers in the aforementioned study focused on developing an established way to resolve the issues with phishing website and guiding the victims to the same site. The researchers focused on analysing 860 phishing emails and 6950 non fishing emails.

In the past, a lot of effort has been put into spam detection (email spam, web spam, SMS spam). Phishing is a very common attack on the subject of email spam. Phishing is the theft of personal information such as login ids, passwords, credit card numbers, and other sensitive information for malevolent purposes. In their research, Fette [40] has demonstrated that phishing attacks may be easily recognized with high accuracy. They used machine learning to analyse user-generated feature sets such as IP-based URLs, the age of domain names connected to them, non-matching URLs, links to non-modal websites, HTML emails, the number of links, domains, dots, and spam- filter output. The proposed technique proved successful in detecting phishing websites and emails that guide victims to those sites. The accuracy of their system was tested on a set of 860 phishing emails and 6950 non-phishing emails, and it was found to be above 96% [40].

Authors_Name	Purpose	Approach
Karim et a.	Spam and Fraud email detection	Revision based Artificial Intelligence and Machine Learning related Approaches
Agarwal and Kumar	Textual Dataset based Email spamming extraction based on detection and creation method	Approaches based on Machine learning algorithms
Harisinghaneyet al. (2014)	Recognising Texts and Figures	The focus has remained on applying K Neural Network Algorithm and the reversal based DBSCAN Algorithm
Mohamad and Selamat	Selecting the Characteristics	Team Based Inverse Document Frequency that focuses on characteristics generation for Rough Pure Mathematics based hybrid selection.

III . Implementation & Methodology:

3.1 Proposed Work

The proposed method of developing the “Naive Bayes accuracy” and the lesser rate of “false positives” are the most useful method for the developing of the processes and the maintenance of security. The algorithms related to the “Machine Learning” that can be predicted within the result of situation related to the machine learning. Despite of the higher accuracy rate, the illegal workers are there to misuse the technical issues associated with algorithms. Developing the protected systems of algorithm, this study, has used the package of “nltk python”, named “stop-words”, has the gathering of various spam words within that. Protecting the algorithms related to the diminishing of accuracy level, these packages are used within the machine learning. Those algorithmic functions help the detectors to investigate easily about the related issues of algorithms. The packages of “stop-words” maintains the updated filed for detecting trashes in electronic messages. Hence it is very significant in investigating the trashes among messages easily and with more efficiency.

The future development of the “proposed works” requires for the enforcement of the extraction of various features. This program me makes this easy to develop the training procedure and the “accuracy” of the “spam detection model.”

Overview of Proposed Work

This research the “e-mail data set” is uploaded from the particular website named “Kaggle” as the method for the purpose of “training dataset.” Due to achieve the better performance level, the embedded datasets have removed the duplicate information and the dropping of those data are usual. After the calculation, the acquired “dataset” is classified into two sets of data including the “train dataset” and “test dataset” within the ratio of “80:20.” Both of these two types of information set are supplied as “text-processing parameters.”

Additionally, all the symbols and the punctuations are removed and those are replaced by the words. During the processing of text, the punctuation symbols and words in the list of stop words are eliminated and replaced with clean words. After that, the “Featured Transform” command is enforced within the process of “clean phrases.” Those words are gathered from the above mentioned “vocabulary” and the transformation of features machine learning. The collected data are enforced in the system of tunning the “hyper parameter” that engages the best values for the identifying the utility based of the sets of information.

Those values are well established with the “random state” in obtaining the values from the process of “hyper parameter tweaking.” The trained model and the features of that model are the process of testing of “unknown data.” The machines are then trained by the values that are gathered by using the systems of the module of “Python Sklearn.” The information related to the process are done with the extraction process that transformed the extraction of various features. The set of information is prepared for the “splitting process.” After that, those set of information are classified into two sets of classifications of “training data set of 80%” and the “testing dataset” with 20% portion.

After the splitting information is gathered, the set of information is diversified with the “Logistic Regression Classifier.” This process initiates the ”accuracy” in acquiring the results in comparing with others. Apart from that, the embedded system can be improved the algorithm with the replacement of the weakest features of different classification.

Varius types of significant enablers are there in predicting the illegal issues within the electronic messages system. At present the uncountable messages and chats are delivered among people and thus investigating the challenges in electronic messages becomes relatively complex and within this periphery of this “restricted corpus.” This study deals with the process of investigating the process of filtering the contents based on the messages rather than the “domain name.”

3.1.1 Logistic Regression factors

These factors are using for distributing the “email spam” and the filtration of them are enabled within the “training data base” to create the useful model and the performance of the models are investigated by using the “tested database.” These types of factors are enforced for the classification and identification of various information.

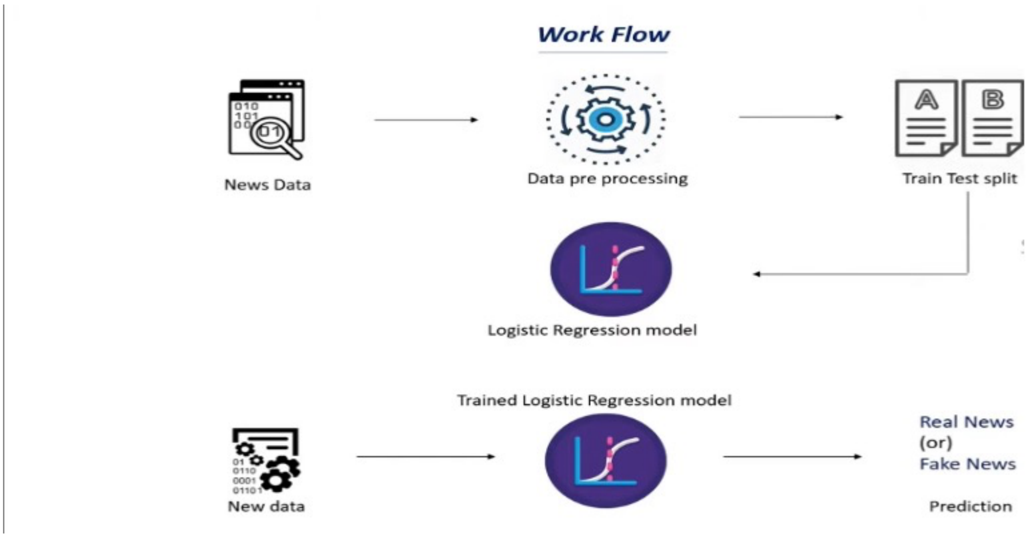


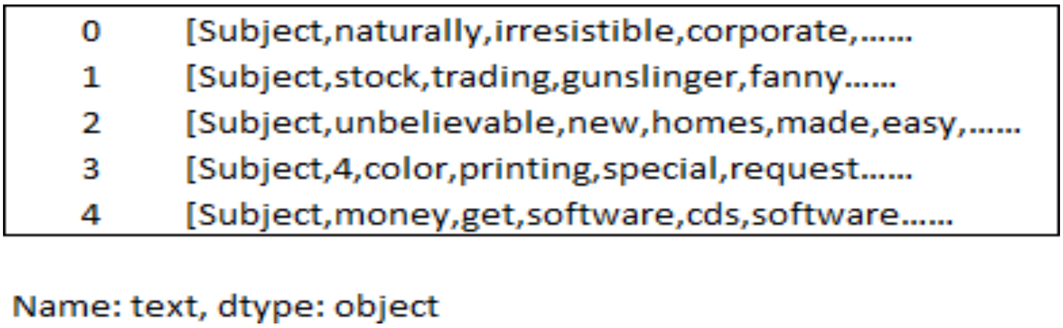
Figure 3.1. Logistic Regression factors

3.2 Pre Processing

Investigating the models and training the raw information are the two ways for “data processing.” The abovementioned process holds the process of “cleaning information” that provides the accuracy and the capacity of the model. After eliminating the “null values” and interchanging them with meaningful values, the process develops the system.

3.2.1 Process of “Tokenization”

“Tokenization” is the most significant way of the “pre-processing stage”. This process engages the acquiring of information and classification of the information into “words and the count” is being counted for each word that presents in the information-set. The use of similar words within the process can be identified by the use of “count vector.” The set of numbers are classified to each word that provides the frequency system. This significant process is described below



```
0       [Subject,naturally,irresistible,corporate,......
1       [Subject,stock,trading,gunslinger,fanny......
2       [Subject,unbelievable,new,homes,made,easy,......
3       [Subject,4,color,printing,special,request......
4       [Subject,money,get,software,cds,software......

Name: text, dtype: object
```

Fig. 3. Preview of Dataset

Reducing the Word Stop

After the information has "tokenized", the process is free from the burden of different words and different punctuations like the "whitespaces, commas, full stops, colons, semi-colons, and other punctuation marks." This process can be drafted using the model named the "NLTK" where the tools comprise various "toolbox" which reduces the unused phrases and symbols, creates the marks of the removing process of stop words.

Originating

This process includes the "stemmed words" that are transformed into the useful format. The variables and the classifications are diversified with the variables on the sets of elementary processes. This process helps in introducing the accuracy in output version of algorithms. Thus, the extraction process that is associated with the system takes the easy process and takes a little time.

Selection and Extraction process

The used format of the "converted process" are transformed into the "manageable format" and the words used within the process of machine learning. This process includes the variables with the "original information" of this process. The extraction process even helps the production of processes with more accuracy and the output related to the process. Research shows that, the more valuable extraction process will be there, the less time will be used to produce the model. Thus, the process of "feature extraction" and the "model creation" are interlinked with each other within the algorithm.

Classification of information

The process of "information classification" can be recognized as the most important process based on the "training and testing" the sets of information that are used within the classification of information used within the model. Normally, the information is classified in various types of forms within the embedded form. This model considered with the usage of 80 % of information for predicting the module and the 20% of information set for evaluating the level of performance as the result of the testing sets of information.

3.3 Processes related to Diversifications

Various diversifications are there that are given for the upper approaches including the "Naive Bayes, Bayesian Network, Support Vector Machine (SVM), Genetic classifier, Random Forest, XGBoost, and Long Short-Term Memory (LSTM)."

3.3.1 Naïve Bayesian Algorithm Application

Depending on the Naive Bayesian Classification Technique, the independent method is developed depending on the assumption prediction conduction. A particular characteristic evaluation is used by this method for classifying depending on presence and absence of other features. The building criteria of Naive Bayesian Network system Method is typically easy and the process is highly useful in understanding the criteria related to large sets of data. Moreover, the simplicity of Naive Bayesian Classification Technique along with the strength of assumptions give rise to an idea that not any single variable is dependent on another one. Class is the sole condition of evaluating these types of independence that includes proper assumption and the most significant benefit of this algorithm is that the required training dataset is comparatively lower than any classifying methods. The case dependent methodology is used by the algorithm and the process speed gets increased properly depending on the singularity of the scams. The resultant variable that is desired to obtain from the evaluation is Posterior Probability named P(c|x) from the three already known variables including P(x), P(c) and P(x|c). The class of the predictor x does not depend on the class of given variable c upon other predictors. Based on the theorem of Naive Bayesian system, the rule for evaluating the Posterior Probability includes:It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier

assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
P(c|x)= P(x|c)XP(c)/P(x)

The Variables are defined as:
P(c|x)= The target class based posterior Probability depending on a given set of attributes
P(x|c)= The Hood based probability of given class predictor
P(c)= The already existing class probability
P(x)= The already existing predictive Probability

3.3.2 Network based on Bayesian Theorem

The scoring function acts as the main classifier in identifying the features of Bayesian network. The connectivities and the orienting strategies are the main factors of structures that get limited depending on the scoring function. The parameters are as well marginalised based on the likelihood of the usability of the required sets of data. The distinction of the class is decided by the Nodes present in the “Markov Blanket” that denotes a conjuncture of the parents, children and the parents of the children. All the characteristics include the attributes that are independent and does not get changed with modification in any other of them. However, the complications related to the Bayesian network is highly problematic regarding the Naive Bayesian network but their levels of performance are almost equal. The reason of the Bayesian network performs worse than the Naive Bayes is that it considers more than 15 cases as attributes. The structure generates the learning ability that discards important isolation criteria based different unimportant attributes. The combined application of Naive Bayes and Bayesian algorithm can help in becoming the augmented Naive Bayes tree and K neural networking Bayesian Framework.

3.3.3 Machines related to the “Support Vector”

A supervised learning algorithm that offers an alternative view of logistic regression, the simplest classification algorithm, is the support vector machines or SVMs. Support vector machines try to find a model that precisely divides the classes with the same amount of margin on either side, where the support vectors are called samples on the margin. Support Vector Machine (SVM) also known as Support Vector Network in machine learning is a supervised learning technique used for classification and regression. In simple terms, given training examples set, each of them marked belonging to one of any number of categories. SVM training algorithm constructs a model that decides and assigns a new example that falls into one category or the other. Hence SVM classifier is represented by a separating hyperplane. Linear and Radial Functional Basis based on Polynomial.

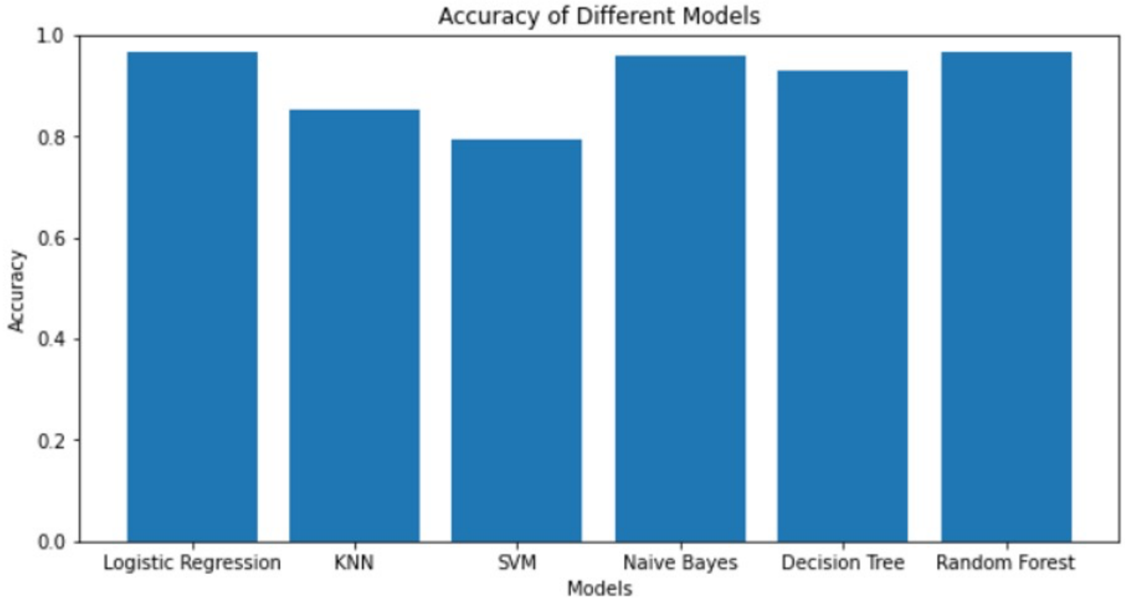
3.3.4 Random Forest Method

Ineffectiveness of decision trees lead to the usage generation for Random Forest Classifier. Random Forest Classifier focuses on generating a bagging based Decision tree generation that focuses on assembling the approach. The trees that are used in this method can be regarded as the most significant approaches bearing ensemble ness within them. The irregular formation of trees that leads to assymetry includes over fitting of training sets. Moreover, Random Forest acts on generating differentiation among various training sets by generating a huge set of different decision trees that are trained for generating higher than average success-oriented decision trees. The transformation of the weak learner to strong ones is conducted easily using this process and the number of solutions is directly proportional to the number of decision trees that are generated.

The Multi Layered Learning of Random Forest method helps in development of individualistic connection generation based on various trees. The votes are created based on overall class divnisioning based on the most number of votes consumed. The building of the decision tree depends on random training subset development using the process of training sets of data. The process of replacement is applied within the entities that gets involved multiple times within a sample and the others do not get appeared before the sample. Random subset of a sample is used depending on the variables and their availability for partition creation depending on dataset nodings. Each decision tree gets increased depending on the noding based quality generation of the partitions. The pruning process is not performed based on the maximum size and the ensemble model does get changed based on Random Forest. However, the resulting decision tree models act as the main identifiers of these Random Forests that is defined by the majority wins gained by each decision tree.

IV Experimental Results

The suggested process after getting implied through the coding of Python, the performance metrics get generated based on employing accuracy. The Precise recalled F1 score is used for accessing the levels of performances.



V. Conclusion:

In this paper, we built a system to help identify spam emails. First, we cleaned up the data by getting rid of unnecessary information and organizing it neatly. Then, we looked at the data to understand it better, like how many spam and non-spam emails there were. Next, we prepared the text in the emails for analysis by breaking it down into smaller parts and removing unimportant words and symbols. After that, we converted the text into numbers so a computer could understand it better.

The Logistic Regression Modelling technique has the highest feasibility among all and the highest accuracy of 0.965 and precision of the recall shows the value of 0.954 that includes F1 score to be 0.938. The accuracy, being high, provides the way to identifying the falsification of positivities along with the negativities in F1 score. The specific use case includes the model performance generation that focuses on problem understanding and specificity evaluation of use case. The model needs high recall value inclusion as a result of performance development based on costing dependency related to filtration capability of the spam emails.

References

[1] CORMACK, G. V. Email spam filtering: a systematic review. Foundations and Trends® in Information Retrieval, 2006, vol. 1, no. 4, pp. 335–455. https://doi.org/10.1561/150 0000006

[2] ZHANG, L., ZHU, J., YAO, T. An evaluation of statistical spam filtering techniques. ACM Transactions on Asian Language Information Processing, 2004, vol. 3, no. 4, pp. 243–269. DOI: 10.1.1.109.7685

[3] DELANY, S. J., BUCKLEY, M., GREENE, D. SMS spam filtering: methods and data.
Expert Systems with Applications, 2012, vol. 39, no. 10, pp. 9899–9908.
DOI: 10.1016/j.eswa.2012.02.053

[4] ZHOU, B., YAO, Y., LUO, J. A three-way decision approach to email spam filtering. In: Canadian Conference on Artificial Intelligence, Lecture Notes in Computer Science, vol. 6085.
Springer, 2010, pp. 28–39. doi: 10.1007/978-3-642-13059-5_6

[5] BARUSHKA, A., HÁJEK, P. Spam filtering using regularized neural networks with rectified linear units. In: Adorni, G., Cagnoni, S., Gori, M., Maratea, M. (eds.) Conference of the Italian Association for Artificial Intelligence. Lecture Notes in Computer Science,
Springer, Cham, 2016, vol. 10037, pp. 65–75. doi: 10.1007/978-3-319-49130-1_6

[6] BHOWMICK, A., HAZARIKA, S. M. E-mail spam filtering: A review of techniques and trends. In: Kalam A, Das S, Sharma K (eds.) Advances in Electronics, Communication, and Computing. Lecture Notes in Electrical Engineering, Springer, Singapore, vol. 443, 2018, pp. 583– 590. doi: 10.1007/978-981-10-4765-7_61

[7] ALMEIDA, T. A., ALMEIDA, J., YAMAKAMI, A. Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers. Journal of Internet Services and Applications, 2011a, vol. 1, vol. 3, pp. 183–200. DOI: 10.1007/s13174-010-0014-7

[8] ALMEIDA, T. A., HIDALGO, J. M. G., YAMAKAMI, A. Contributions to the study of SMS spam filtering: new collection and results. In: Proceedings of the 11th ACM

[9] CHOUDHARY, N., JAIN, A. K. Towards filtering of SMS spam messages using machine Learning-Based Technique. In: Singh, D., Raman, B., Luhach, A., Lingras, P. (eds.) Advanced Informatics for Computing Research. Communications in Computer and Information Science, vol. 712, Springer, Singapore, 2017, pp. 18–30. doi: 10.1007/978-981-10-5780-9_2

[10] KAUR, R., SINGH, S., KUMAR, H. Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches. Journal of Network and Computer Applications, 2018, vol. 112, pp. 53–88. doi: 10.1016/j.jnca.2018.03.015

[11] CRAWFORD, M., KHOSHGOFTAAR, T.M., PRUSA, J.D., RICHTER, A.N., AND AL NAJADA, H. Survey of review spam detection using machine learning techniques.

Journal of Big Data, 2015, vol. 2, no. 1, pp. 1-23. DOI: 10.1186/s40537-015-0029-9

[12] LE, Q., MIKOLOV., T. Distributed representations of sentences and documents. In: International Conference on Machine Learning, 2014, vol. 32, pp. 1188–1196.

[13] LECUN, Y., BOTTOU, L., BENGIO, Y., HAFFNER, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, vol. 86, no. 11, pp. 2278–2324. https://doi.org/10.1109/5.726791

[14] LILLEBERG, J., ZHU, Y., AND ZHANG, Y. Support vector machines and word2vec for text classification with semantic features. In: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2015, pp. 136–140. doi: 10.1109/ICCI-CC.2015.7259377

[15] HOANCA, B. How good are our weapons in the spam wars? IEEE Technology and Society Magazine, 2006, vol. 25, no. 1, pp. 22–30. DOI: 10.1109/MTAS. 2006.1607720

[16] LAORDEN, C., UGARTE-PEDRERO, X., SANTOS, I., SANZ, B., NIEVES, J., BRINGAS, P. G. Study on the effectiveness of anomaly detection for spam filtering.

Information Sciences, 2014, vol. 277, pp. 421–444. DOI : 10.1016/j.ins. 2014.02.114

[17] OBIED, A., ALHAJJ, R. Fraudulent and malicious sites on the web. Applied Intelligence, 2009, vol. 30, no. 2, pp. 112–120. https://doi.org/10.1007/s10489- 007-0102-y

[18] WEI, C. P., CHEN, H. C., CHENG, T. H. Effective spam filtering: a single class learning and ensemble approach. Decision Support Systems, 2008, vol. 45, no. 3, pp. 491–503. DOI: 10.1016/j.dss.2007.06.010

[19] NEXGATE. 2013 State of Social Media Spam Research Report, 2013. Accessed 10 January 2020, available at: https://go.proofpoint.com/rs/309-RHV-619/images/Nexgate- 2013- State-of-Social-Media-Spam-Research-Report.pdf.