SURVEY ON IDENTIFICATION OF SPAM MESSAGES, WARNING AND BLOCKING THE SENDER

Dr. K M Shivaprasad[1] ,SreeRakshita Lahari[2] ,Pallavi D[3],Saragu Shreevani[4],Sahanashree TG[5]
shivakalmutt@gmail.com[1],lahari.cse.rymec@gmail.com[2],pallavid.cse.rymec@gmail.com[3],
saragushreevani.cse.rymec@gmail.com[4],sahanashreetg.cse.rymec@gmail.com[5]
Department of Computer Science and Engineering
Rao Bahadur Y Mahabaleshwarappa Engineering College, Ballari, Karnataka, India
Affliated to Visvesvaraya Technological University, Belagavi

1.ABSTRACT

The messages are still the primary choice as communication medium. Nowadays the spam messages are major issue in the society, there are different types of fraud messages which are affecting and diverting the society to the negative mode by various threats.In recent years, various researches have come across with detection and identification of spam messages by classifying the messages as positive or negative by using traditional classification algorithms, this has been carried out in order to classify and cluster the messages into different groups but they have not carried out any certain tasks like warning the spam message sender and blocking the messenger.

In our paper, we propose a new algorithm which can help us in order to identify whether the messages sent by the messenger is spam or not and perform the task of warning the messenger if the message sent by them is found inappropriate and blocking the messenger if required by fixing threshold value. This proposed methodology may provide a very good accurate result compared to different traditional algorithms.

Keywords:Logistic Regression (LR),Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB) and Decision Tree (DT).

2.INTRODUCTION

The Short Message Service (SMS) has been widely used as a communication tool over the past few decades as the popularity of mobile phone and mobile network grows.SMS Spanning has become a major nuisance to the mobile subscribers. Subscribers given its pervasive nature. The SMS spam also known as Drunk Messages, refers to any irrelevant messages delivered using mobile networks. There are several reasons that lead to the popularity of spam messages. There are large number of users who use mobile phone in the world, making the potential victims of spam messages attacks also high. The capability of spam classifier on most mobile phones is relatively weak due to the shortage of computational resources, which limits them from identifying the spam message correctly and efficiently

Identification and classification of spam messages is one of the trending era in the field of Artificial Intelligence with Machine Learning Algorithms. In our paper we are using Multinominal distribution theory which includes Naive Bayes Classifier which help to identify and classify the text as spam or not spam. We are using the text SMS data set which contains the 5000 samples which is included with spam and not spam text message.

In today’s digital era the digital messages are playing vital role in field of text representation.

Themessages are still the primary choice as communication medium.Nowadays the spam messages are major issue in the society, there are different types of fraud messages which are affecting and diverting the society to the negative mode by various threats, in order to overcome.

In recent years, various researches have come across with detection and identification of spam messages by classifying the messages as positive or negative by using traditional classification algorithms, this has been carried out in order to classify and cluster the messages into different groups but they have not carried out any certain task like warning the spam message sender and blocking the messenger. This problem we have chosen ahybrid algorithm to identify whether the message sent by the sender is spam or not.

In our paper, we propose a new algorithm which can help us in order to identify with the message sent by the messenger is spam or not and perform the task of warning the messenger if the message sent by them is found inappropriate and blocking the messenger if required by fixing threshold value. This proposed methodology may provide an accurate result compared to different traditional algorithms

In this paper we contribute to the following,

(i) This study discusses various machine learning based spam filters, their architecture, along with their pros and cons. We also discussed the basic features of spam email.

(ii) Some exciting research gaps were found in the spam detection and filtering domain by conducting a comprehensive survey of the proposed techniques and spam's nature.

(iii) Open research problems and future research directions are discussed to enhance email security and filtration of spam emails by using machine learning methods.

(iv) Several challenges currently faced by spam filtering models and the effects of those challenges on the models' efficiency are discussed in this study.

(v) A comprehensive comparison of machine learning techniques and concepts that help understand machine learning's role in spam detection is provided.

(vi) The study categorizes different spam detection methods according to machine learning techniques to better understand concepts jointly.

(vii) Various future spam detection and filtration directions are discussed that could be explored to detect spam better and add more security to email platforms.

3.Litrature Survey

In an attempt to alert users against fake messages and mischief, our paper has presented a study to automatically analyze the information credibility of messages propagated [7].

Ammar et al used factual classifiers in the spam channel to find out the calculated the false positive and false negative Rates by calculating plashing index and ham messages [11].

There is rapid increase in the interest being shown by the global research community on the spam messages detecting. In this section, we present similar scenarios that have been presented in the literature in this domain. This method is followed so as to articulate this issues that are yet to be addressed and highlight the difference with our current review [7].Spam messages are any type of unwanted or harmful messages, such as advertisements frauds, business services etc .They annoy end users, consume the research of mobile devices, involving memory spaces and lead to overlooking message channels [8]. This approach is followed in order to point out what issues have yet to be addressed and to highlight the difference with our current analysis .For the preliminary benchmark experiment, here the fingerprint of every newly received message to the fingerprints of all identified spam texts. Those related to any of the already identified spam fingerprints were classified as spam [12].

Based on last number of messages users, the short messages center receives a massive amount of messages stream at a very high speed spam messages only make things worse, thus the current index design is somehow random until the users themselves classify their perspective of wanted v/s unwanted messages[9].There are several different machine learning based classification applications proposed in the field of messages spam detection, a greater number of these approaches are based on traditional machine learning techniques such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Naïve Baye's (NB) and Decision Tree (DT) [10].

In Jain et al proposed a method to apply rule-based case-based models on message spam detection problem. The authors extracted rules and implemented Decision Tree (DT), RIPPER and PRISM and the DT, yielding 99.01% Tree negative Rate (TNR) and 92.82% Tree positive Rate (TDR) [10].Before we exploit the spam depending, we first investigate existing spam detection and warning techniques. There are broad range of solutions to mitigate the spam problem.

- Black/White List: A general list-based spam blocking technique that limited messages from black list (Known spammer addresses or malicious mail server IP address) and only allows messages from the white list.

- Content based filtering: simply spots message that contains some keywords that are commonly used in spam.
- Response and challenge based: It tries to shift part of the responsibility to the sender.
- Collaborate and spam filtering collect spam information based on the user feedback.
- Blocking incoming messages from unknown senders based on white list.
- User blocks spammer based on the black list.
- Limit the automate IM user registration by including registration process [8].

To develop the efficient spam solution, it has to support important features which are listed as follows:

- Real-time filtering support spam messages should make decision on real time in-order for it to be useful for mobile device uses.
- Self-learning spam filtering and identifying include machine learning algorithm to adapt a new kind of spam as spammers are changing their ways constantly.
- Privacy considers as a type of private communications and should not be vulnerable to any type of misuse by third-parties.
- Black and white personalization- spam message classification different from one person to other system should adopt to user preferences [11].

In any research area, there is always place for improvement. Many researchers have suggested new ways to integrate machine learning techniques into spam filter and blocking in the hope to achieve better performance [9].Recently, machine learning algorithms have been introduced in research paper as a solution to this problem. Researches like Sethi et al have contributed to comparing the performance of different machine learning algorithms. They have tested Naïve Bayes, Random Forest & logistic regression & concluded that Naïve Bayes performed the best among others. Gupta have compared 8 different classifiers, their comparison concluded that convolutional neural network classifier scores the highest accuracy 98.25%, 99.19%.

Healy et al have compared 3 different classifiers, which are K-nearest neighbor, support vector machine and naïve based. These research papers have conducted a comparative survey between unsupervised algorithms used for automation, classification and maintenance. In comparison,

they have considered against 16 parameters such as performance, accuracy, robustness, complexity, reliability etc. [9].

Bovjnoum have proposed a message spam filter and warning based on text messages, enhanced version of SVM. His experiment result has shown that choosing optimal value for suggested filter parameters it reaches an accuracy of 95.13% in training set and 89.32% in testing test. It has successfully proven to be suitable model for text messaging classification [9].

The goal of this survey is to undertake a thorough literature evaluation on approaches for identifying, detecting, and blocking the spammer and the spam warning content in messages. Our efforts are primarily motivated by a desire to learn more about difficult spam text detection and categorized algorithms. This section discusses the survey methodology that we used to conduct our detailed spam detection and blocking review [12].Following up from the previously mentioned contributions, this paper will contribute to the field of knowledge by introducing a new hybrid system that includes the technique based on unsupervised machine learning. As spammers continuously changing their style of writing spam, the benchmark in the proposed system is that it will achieve higher accuracy by distinguishing both trained spam filters and the unknown ones [9].

4.Background Knowledge

4.1Methods

4.1.1. Standard Spam Filtering Method.

Standard spam filtering is a filtering system that implements a set of rules and works with that set of protocols as a classifier. At first, content filters are implemented and use artificial intelligence techniques to figure out the spam messages.

The message header filter, which extracts the header information from the message, is implemented in the second step. After that, backlist filters are applied to the messages to clinch the messages coming from the backlist file to avoid spam messages. After this stage, rule-based filters are implemented, recognizing the sender using the subject line and user-defined parameters. Eventually, allowance and task filters are used by implementing a method that allows the account holder to send the message [7] as shown in fig 4.1.1.

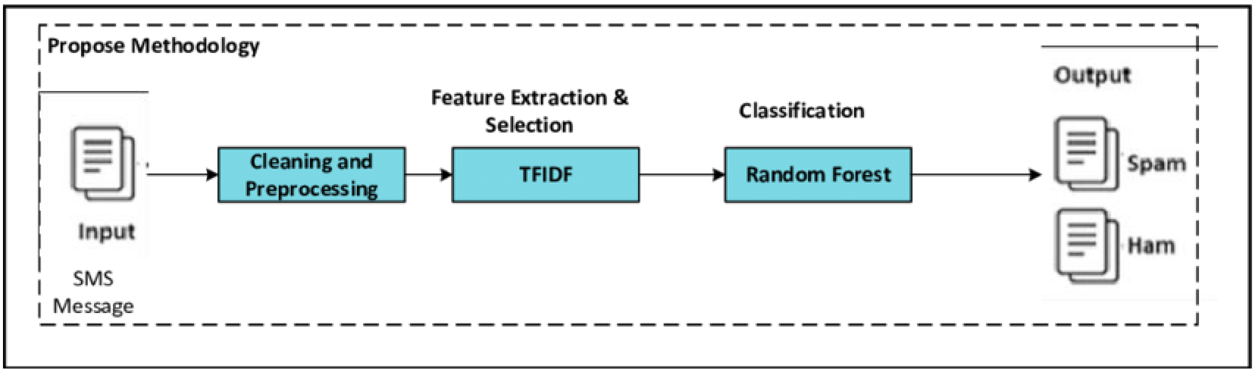


Fig 4.1.1 Standard spam Method

4.1.2.Enterprise Level Spam Filtering Method.

Message spam detection at the enterprise level is a technique in which various filtering frameworks are installed on the server, dealing with the message transfer agent and classifying the collected messages into one spam or ham. This system client uses the system consistently and effectively on a network with an enterprise filtering technique to filter the messages [8].

Existing methods of spam detection use the rule of ranking the message. A ranking function is specified in this principle, and a score is generated against every post. The junk message or ham message is given specific scores or ranks as shown in the fig 4.1.2. Since spammers use different approaches, all tasks are regularly modified by implementing a list-based technique to block the messages automatically. The below is reproduced from Bhuiyan et.it shows the architecture of the client and enterprise level spam filtering process [9][10].

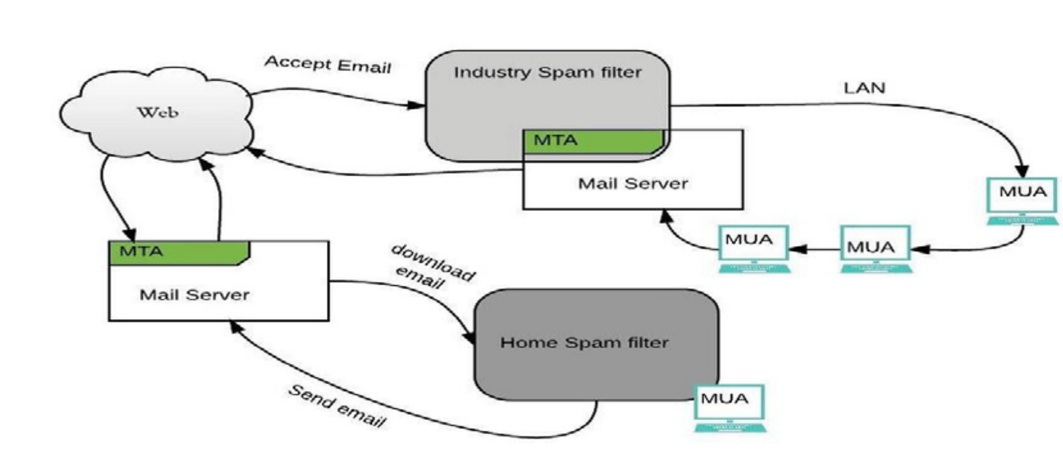


Fig 4.1.2 Enterprise Level Spam Filtering Method

4.1.3. Case-based spam filtering Method.

One of the well-known and conventional machine learning methods for spam detection is the case-based or sample-based spam filtering system [11].

As shown in the fig 4.1.3, there are many phases to this type of filtering with the aid of the collection method; it collects data during the first step [11].

Finally, the machine learning technique is extended to training sets and test sets to determine whether this is a message. The final decision is made through two steps self-observation and classifier’s result, deciding whether the message is spam or legitimate [12].

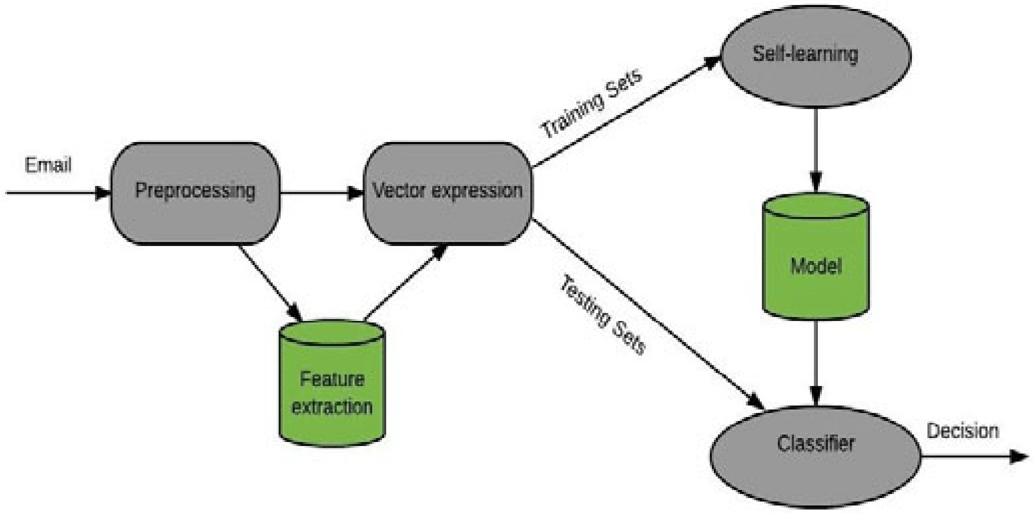


Fig 4.1.3 Case-based spam filtering method

4.1.4.Supervised Machine Learning Method.

Supervised machine learning algorithms are machine learning models that need labeled data. Initially, labeled training data is provided to these models for training, and after training models predict future events. In other words, these models begin with the analysis of an existing training dataset, and they generate a method to make predictions of success values. Upon proper training, the system can provide the prediction on any new data related to the user's data at the training time. Furthermore, the learning algorithm accurately compares the output to the expected output and identifies errors to modify the model[13].

Supervised learning uses labeled data for training as shown in the below fig 4.1.4, and then it can predict the new data. This type of learning can be used in solving various problems, i.e., advertisement popularity, spam classification, face recognition, and object classification [13][14].

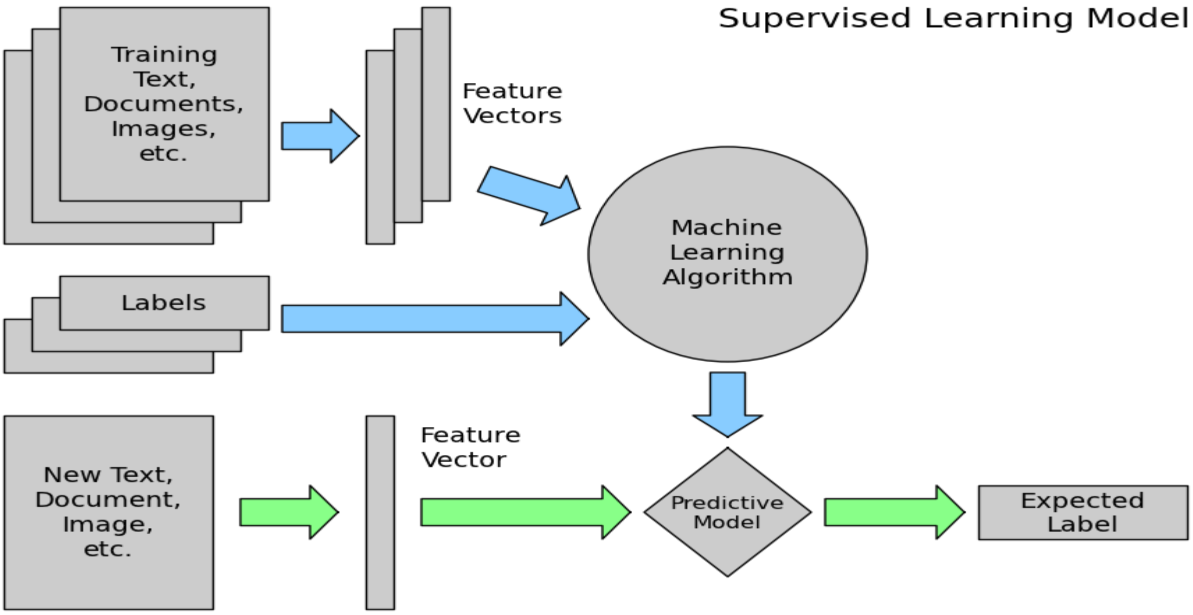

Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Labels
Machine Learning Algorithm
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Fig 4.1.4 Supervised Machine Learning Method

4.1.5. Decision Tree Classifier Method

Decision tree classifier is a machine learning algorithm, which has been widely used since the last decade for classification as per below fig 4.1.5. This algorithm applies a simple method of solving any problem of classification. A decision tree classifier is a collection of well-defined questions about test record attributes. Each time we get an answer, a follow up question is raised until a decision is not made on the record [15].

Tree-based decision algorithms define models that are constructed iteratively or recurrently based on the data provided. The decision tree-based algorithms goal is used to predict a target variable's value on a given set of input values [16].

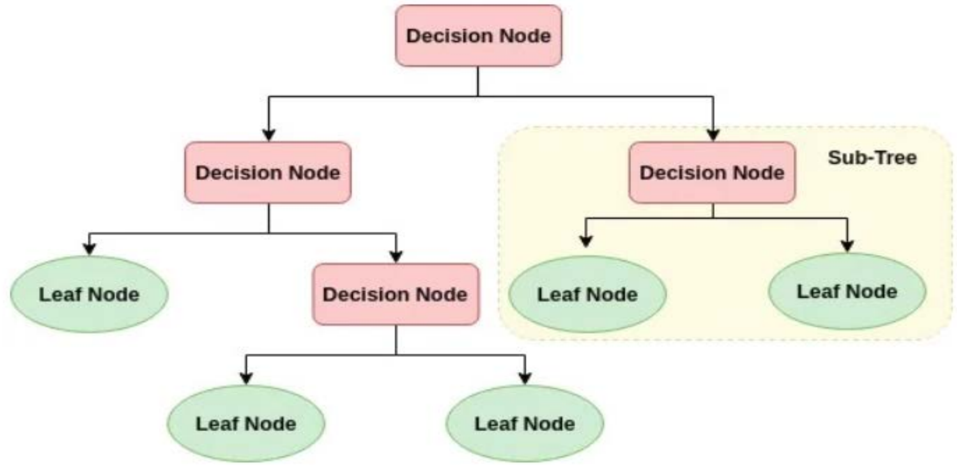


Fig 4.1.5 Decision Tree Classifier Method

4.1.6. Support Vector Machine (SVM) Method.

The support vector machine (SVM) is an essential and valuable machine learning model. SVM is a formally defined discriminative supervised learning classifier that takes labeled examples for training and gives a hyperplane as output, classifying new data as shown in below fig 4.1.6.

SVM got an accuracy of 94.06% in his work, and the extreme learning machine (ELM) model got a 93.04% accuracy level, suggesting just 1.1% performance improvement that SVM achieved over ELM. He indicated that SVM's improvement over ELM accuracy is marginal. It implies that, in situations where detection time is critical, as in real-time systems, the ELM spam detector should be given preference over SVM spam detection. Although SVM got a higher accuracy level in his research, it takes more time for training than the ELM system [17].

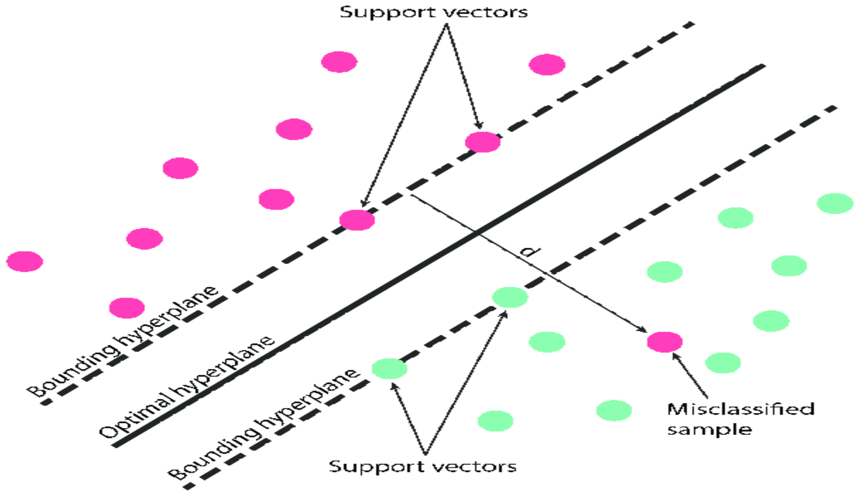


Fig.4.1.6 Support Vector Machine (SVM) Method

4.2Techniques

4.2.1.Naïve Bayes Classifier Technique.

The Naïve Bayes classifier [18] is based on the Bayes theorem. It assumes that the predictors are independent, which means that knowing the value of one attribute impacts any other attribute's value as given in the fig 4.2.1. Naïve Bayes classifiers are easy to build because they do not require any iterative process and they perform very efficiently on large datasets with a handsome level of accuracy. Despite its simplicity, Naïve Bayes is known to have often outperformed other classification methods in various problems.

Naïve Bayes uses probability for classification, and the probability is counting the frequency and combination of values in a dataset. This research uses three steps for the filtration of messages, i.e., preprocessing, feature selection, and, at last, it implements the features by using the Naïve Bayes classifier. The preprocessing step removes all conjunction words, articles, and stop words from the message body [19].

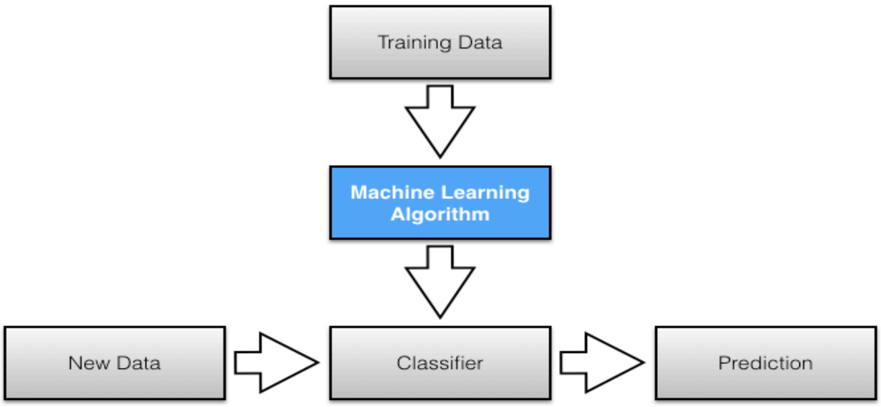


Fig 4.2.1Naïve Bayes Classifier Technique.

4.2.2.Artificial Neural Networks Technique.

An artificial neural network (ANN) is a computational model based on the functional aspects of biological neural networks, also known as the neural network (NN)[20]. Many sets of neurons are joined in a neural network, and information is interpreted using a computational approach connection. In most situations, an ANN is an adaptive system, which changes its structure depending on external or internal information flowing through the network during the learning phase. Current neural networks are nonlinear approaches to statistical data processing. These are commonly used when there are complex relationships between inputs and outputs or unusual performance patterns. In the fig 4.2.2, this method is used for the detection of spam in online social networks.

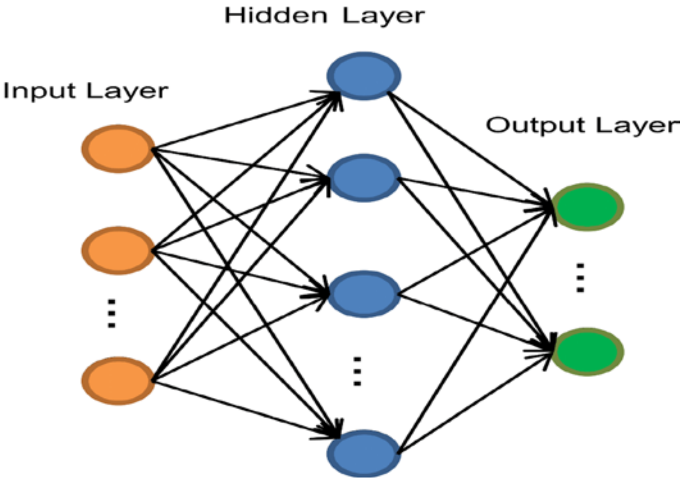


Fig 4.2.2 Artificial Neural Networks Technique

4.2.3. Hierarchical Clustering Technique.

Hierarchical clustering identifies clusters with a hierarchy achieved either by iteratively combining smaller clusters into a more significant cluster or by splitting a more massive cluster into smaller clusters. This cluster hierarchy, generated through a clustering algorithm, is called a dendrogram. A dendrogram is one way of representing the hierarchical clusters as shown in the below fig

4.2.3. The user can understand different clusters based on the level at which the dendrogram is defined. It uses a similarity scale representing the distance between the clusters grouped from the massive cluster [21].

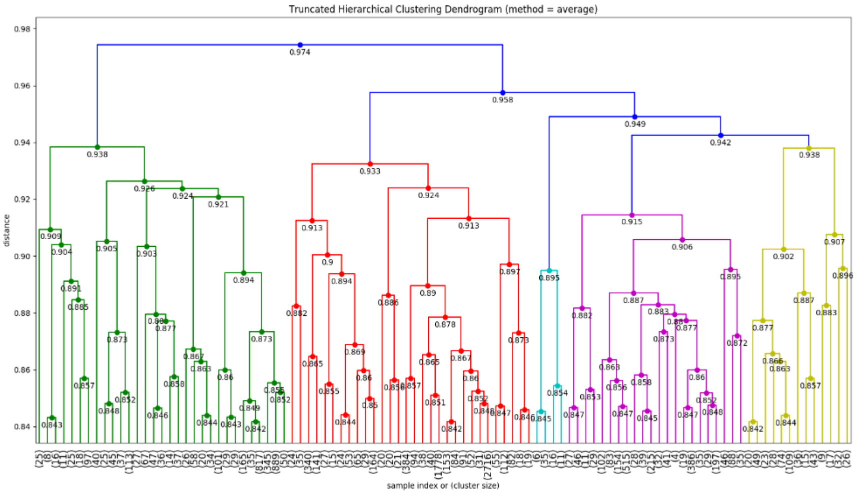


Fig 4.2.3 Hierarchical Clustering Technique

4.2.4. Partitional Clustering Technique.

A partitional clustering divides a single set of data objects into nonoverlapping subsets (clusters) so that each data object is in only one subset. Partitional clustering algorithms make different partitions of data and then evaluate the required results based on some criteria [22].

The partitioning technique forms different partitions of data by using the formula K,each partition represents a cluster based on a set of N points in the data, that is, by fulfilling the following conditions:

(1) Each class contains one point or more

(2) Each point comes as part of exactly one group

4.2.5.Convolutional Neural Network (CNN) Technique.

A convolutional neural network (CNN) is a particular type of artificial neural network which uses perceptron for supervised learning. As shown in the fig 4.2.4 the supervised learning is used to analyze data.There are wide range of applications involved in CNNs. Traditionally used for image processing. CNNs are nowadays used for natural language processing as well. A CNN in relative terms is known as a ConvNet. Similar to other ANNs, a CNN also has an input layer, some hidden layers and an output layer, but it is not fully connected. Some layers are convolutional, that use a mathematical model to pass on the results to layers ahead in the network.

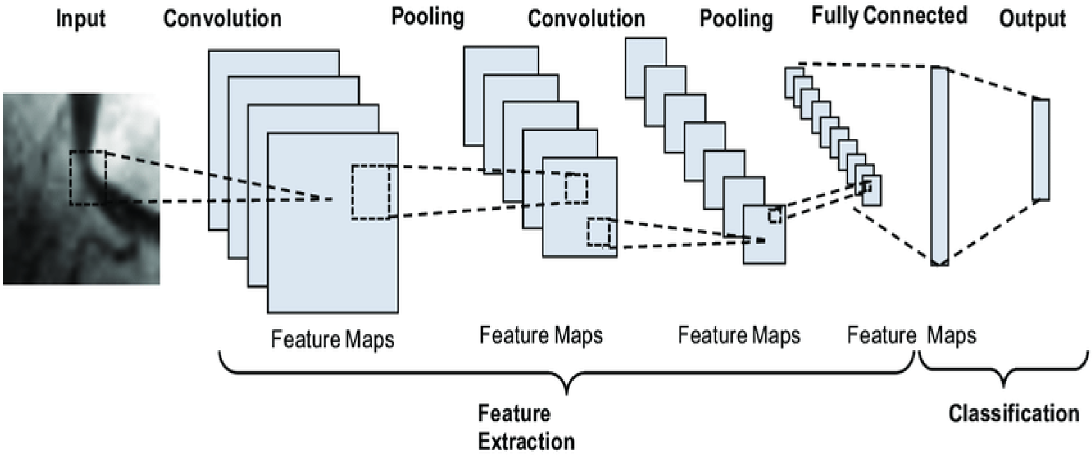


Fig 4.2.4 Convolutional Neural Network (CNN) Technique

4.3ALGORITHMS

4.3.1.Randomforest Algorithm

It is a supervised machine learning algorithm that is used widely in problems such as classification and regression problems, also it can build different samples and decision trees.

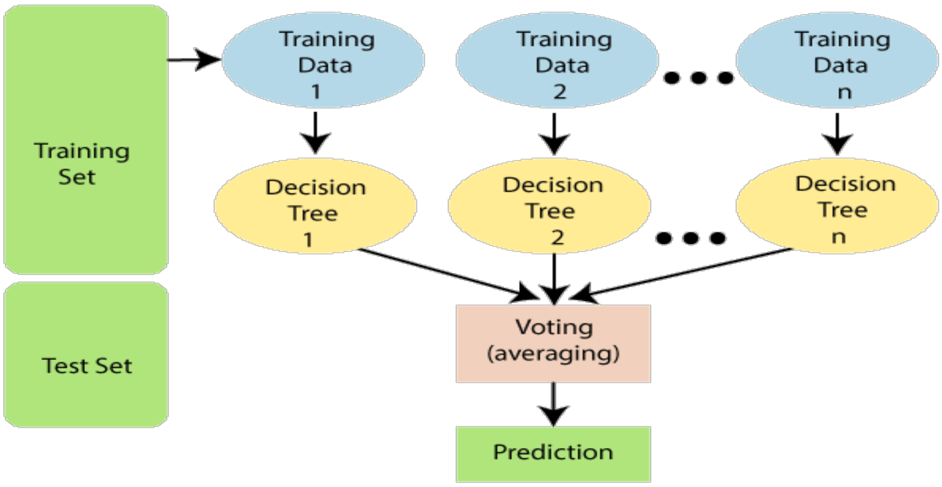
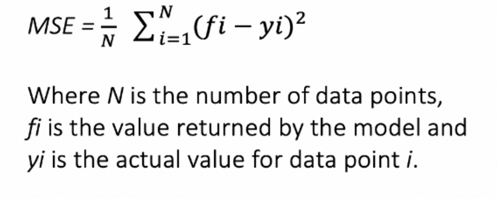


Fig 4.3.1 Randomforest Algorithm

It contains a n number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy as shown in above figure 4.3.1.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The formula 4.3.1 is used for random forest



$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

Where *N* is the number of data points, *fi* is the value returned by the model and *yi* is the actual value for data point *i*.

Formula 4.3.1

4.3.2.Bernoulli Naïve Bayes algorithm

Bernoulli Naive Bayes is a part of the Naive Bayes family. It is based on the Bernoulli Distribution and accepts only binary values either 0 or 1 as shown in below fig 4.3.2. If the features of the dataset are binary, then we can assume that Bernoulli Naive Bayes is the algorithm to be used.

The formula 4.3.2 is used for Bernoulli algorithm

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Formula 4.3.2

where,
P(A|B) = Probability of A given evidence B has already occurred

P(B|A) = Probability of B given evidence A has already occurred

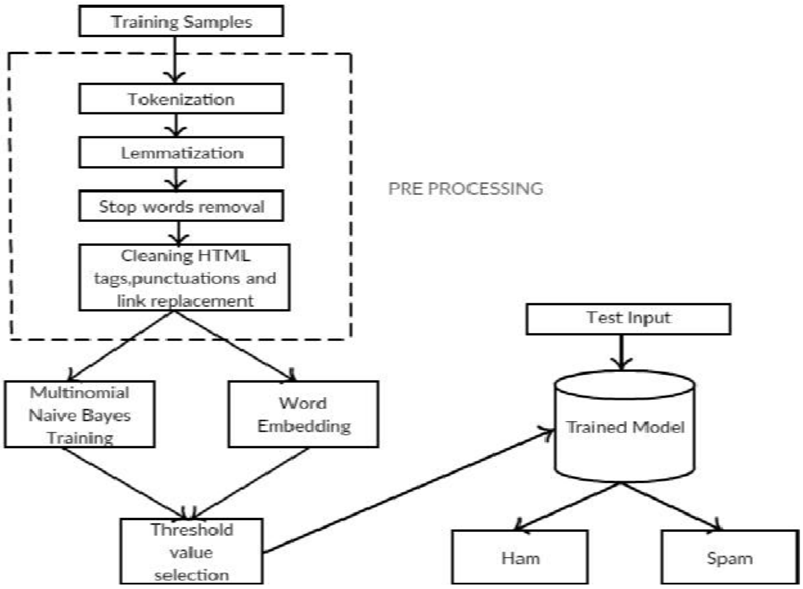
P(A) = Probability that A will occur

P(B) = Probability that B will occur

Multinominal Naïve bayes algorithm

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing. Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features.The Naive Bayes method is a strong tool for analyzing text input and solving problems with numerous classes. Because the Naive Bayes theorem is based on the Bayes theorem, it is necessary to first comprehend the Bayes theorem notion.

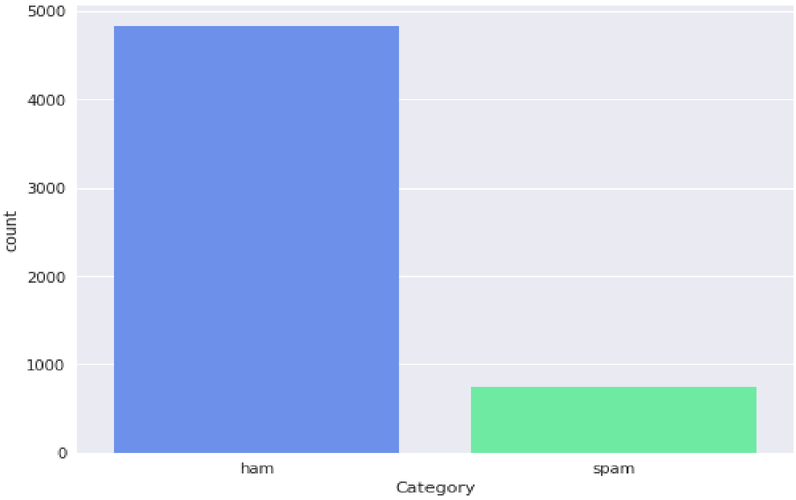
The formula used is P(c|x) = P(x|c) * P(c) / P(x)

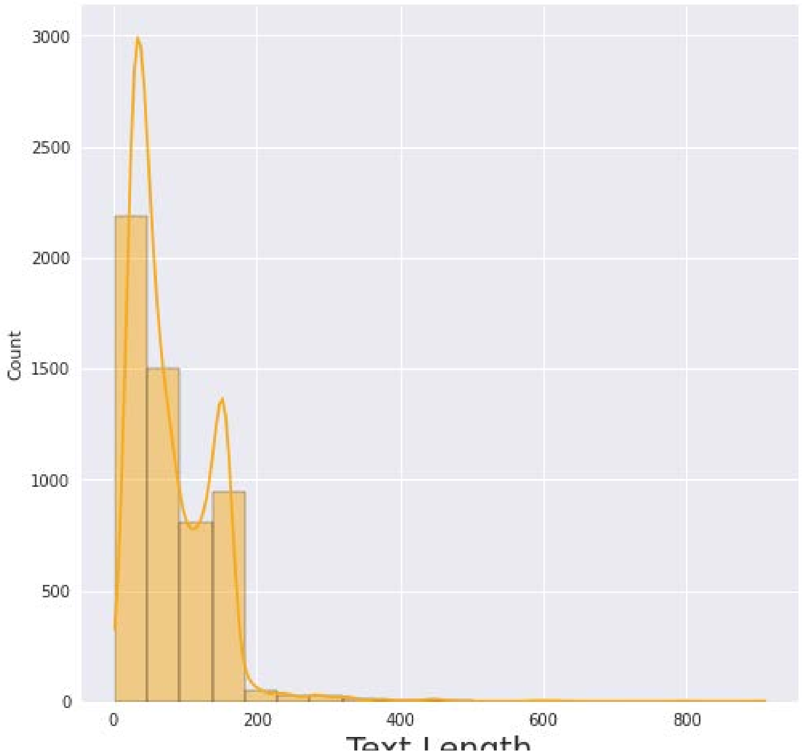


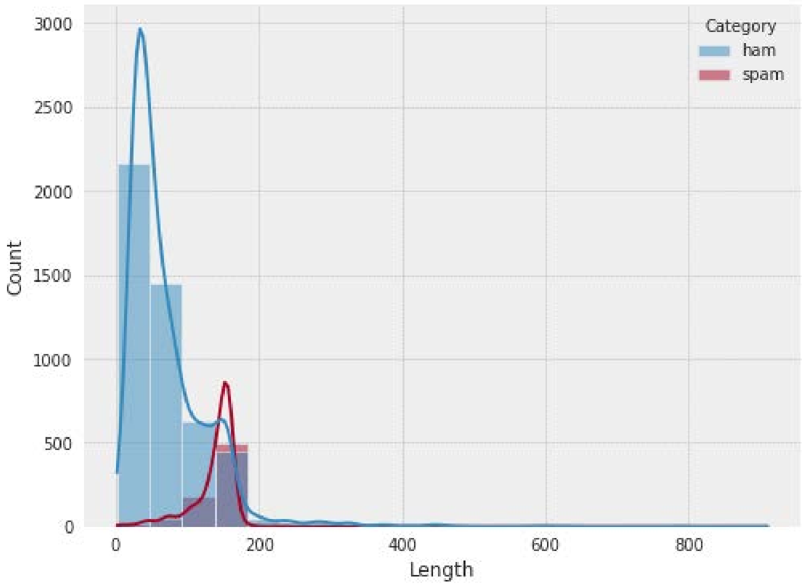
RESULTS AND DISCUSSION:

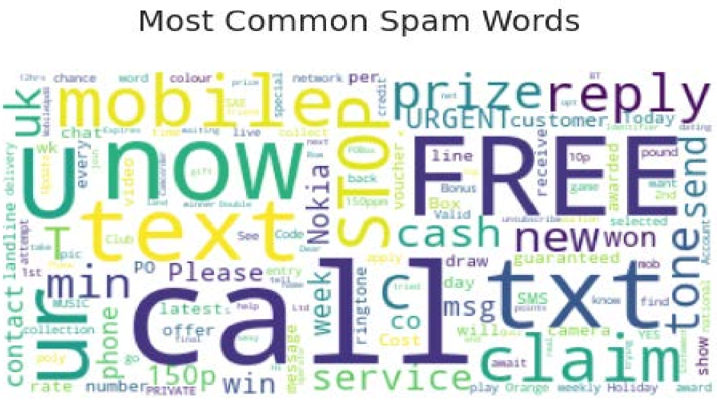
Machine learning algorithms have been extensively applied in this field of spam. Substantial work has been done to improve the effectiveness of spam filters for classifying messages as either ham(valid messages) or spam(unwanted messages) by means of ML classifiers. They have the ability to recognize distinctive characteristics of content of messages. Many significant works have been done in the field of spam detecting and filtering using techniques that does not posses the ability to adapt to different conditions, and on problems that are exclusive to some fields.

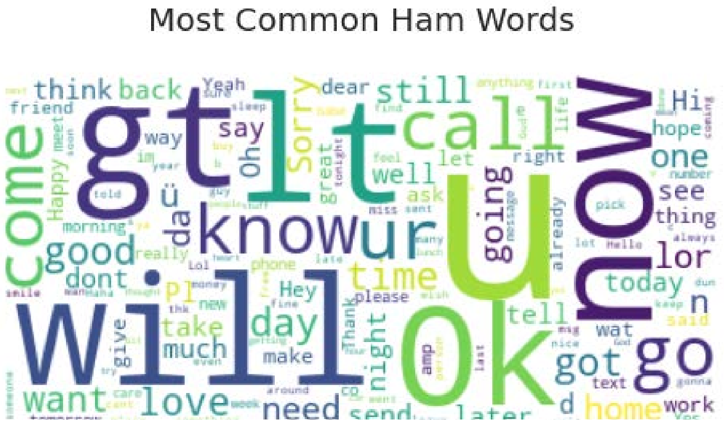
Further research work needs to be conducted to tackle the fact that message spam is a concept drift problem. As such, while the spam filters researchers are trying to increase the prognostic accuracy of filter, the spammers also evolving and trying to surpass the efficiency of the spam filters. It becomes very important to develop more efficient techniques that will adequately handle the trend or progression in spam features to evade many spam filters undetected. The most successful technique applied in the filtering spam is the content-based approach which classifies message as either spam or ham depending on the data that made up content of the message.




Count
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
Text Length


Category
ham
spam
Count
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
Length


Most Common Spam Words
call
FREE
now
txt
u
text
mobile
prize
reply
claim
ur
STOP
cash
new
tone
send
uk
min
Please
service
won
Nokia
win
150p
phone
contact
week
latest
offer
msg
URGENT
customer
today
line
ringtone
awarded
guaranteed
draw
receive
voucher
landline
collection
rate
number
PRIVATE
co
Cost
SMS
camera
pound
day
show
award
weekly
Holiday
Orange
play
await
chat
video
every
wk
chance
word
colour
network
per
special
credit
live
collect



CONCLUSION

In this paper, we have proposed a modified algorithm that aims to identify SMS spam. The main purpose of this paper is to overcome the problems faced by people/mobile users by fraud messages. We have concentrated on the literature review of various papers in order to find out the proper model for solving our problem statement. The expected experiment result might provide best accuracy compared to the other traditional algorithm.

REFERENCES

[1] P.K.Roy, J.P.Singh, and S. Benerjee, “Deep learning to filter SMS spam,” Future Gener.Comput. Syst., vol. 102, pp. 524-533, Jan. 2020.

[2] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, “A comparative study of spam SMS detection using machine learning classifiers,” inProc. 11th Int. Conf. Contemp. Comput. (IC3), Aug. 2018,pp. 1-7.

[3] A. Ghourabi, M.A. Mahmood, and Q.M. Alzubi, “A hybrid CNN-LSTM model for SMS spam detection in Arabic and English messages,” Future Internet, vol. 12, no. 9, p. 156, Sep. 2020.

[4] S. Mishra and D. Soni, “Smishing detector: A security model to detect Smishing through SMS content analysis and URL behaviour analysis,” Future GenerComput. Syst., vol. 108,pp 803-815, jul. 2020.

[5] E. S. D. Reis, C. A. D. Costa, D. E. D. Silveira, R. S. Bavaresco, R. D. R. Righi, J. L. V. Barbosa, R. S. Antunes, M. M. Gomes, and G. Federizzi, “Transformers aftermath.” Commun. ACM, vol. 64, no. 4, pp. 154-163, Apr. 2021.

[6] M. Honnibal, I. Montani, S. Van Landeghem , and A. Boyd, “spaCy:Industrial - strength natural language processing in python,” 2020, doi: 10.5281/zenodo.1212303

[7] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, and J. D. Tygar, “Characterizing botnets from emessage spam records,” LEET, vol. 8, pp. 1–9, 2008.

[8] A. K. Jain and B. B. Gupta, “Towards detection of phishing websites on client-side using machine learning based approach,” Telecommunication Systems, vol. 68, no. 4, pp. 687–700, 2018.

[9] M. F. N. K. Pathan and V. Kamble, “A review various techniques for content based spam filtering,” Engineering and Technology, vol. 4, 2018.

[10] A. K. Jain and B. B. Gupta, “A novel approach to protect against phishing attacks at client side using auto-updated white-list,” EURASIP Journal on Information Security, vol. 2016, no. 1, p. 9, 2016.

[11] A. Bhowmick and S. M. Hazarika, “Machine learning for E-mail spam filtering: review, techniques and trends,” 2016, https://www.researchgate.net/publication/303812063_Machine_Learning_for_mail_Spam_Filtering_ReviewTechniques_and_Trends.

[12] M. Bassiouni, M. Ali, and E. A. El-Dahshan, “Ham and spam e-mails classification using machine learning techniques,” Journal of Applied Security Research, vol. 13, no. 3, pp. 315– 331, 2018.

[13] J. R. M´endez, T. R. Cotos-Yañez, and D. Ruano-Ord´as, “A new semantic-based feature selection method for spam filtering,” Applied Soft Computing, vol. 76, pp. 89–104, 2019.

[14] R. Alguliyev and S. Nazirova, “Two approaches on implementation of CBR and CRM technologies to the spam filtering problem,” Journal of Information Security, vol. 3, no. 1, Article ID 16724, 2012.

[15] J. Tanha, M. van Someren, and H. Afsarmanesh, “Semi-supervised self-training for decision tree classifiers,” International Journal of Machine Learning and Cybernetics, vol. 8, no. 1, pp. 355–370, 2017.

[16] A. Subasi, S. Alzahrani, A. Aljuhani, and M. Aljedani, “Comparison of decision tree algorithms for spam E-mail filtering,” in Proceedings of the 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), IEEE, Riyadh, Saudi Arabia, April 2018.

[17] A. Singh, N. (akur, and A. Sharma, “A review of supervised machine learning algorithms,” in Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, New Delhi, India, March 2016.

[18] I. Rish, “An empirical study of the naive Bayes classifier,” in IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, University of British Columbia, Computer Science Department, Vancouver, Canada, 2001.

[19] N. F. Rusland, N. Wahid, S. Kasim, and H. Hafit, “Analysis of Na¨ıve Bayes algorithm for email spam filtering across multiple datasets,” in Proceedings of the IOP Conference Series: Materials Science and Engineering, IOP Publishing, Busan, Republic of Korea, 2017.

[20] N. Sutta, Z. Liu, and X. Zhang, “A study of machine learning algorithms on email spam classification,” in Proceedings of the 35th International Conference, ISC High Performance 2020, vol. 69, pp. 170–179, Frankfurt, Germany, 2020.

[21] R. Ahuja, A. Chug, S. Gupta, P. Ahuja, and S. Kohli, “Classification and clustering algorithms of machine learning with their applications,” in Nature-Inspired Computation in Data Mining and Machine Learning, pp. 225–248, Springer, Cham, Switzerland, 2020.

[22] W.-F. Hsiao and T.-M. Chang, “An incremental cluster-based approach to spam filtering,” Expert Systems with Applications , vol. 34, no. 3, pp. 1599–1608, 2008.