Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Online Correspondence Hard Sell Ranking Using Expert System and Development of 'Thinking' Computer Systems

Shaik Heena[1], T Jaya sri[2], V Ramya[3], Shaik Sheema[4], D Karunamma[5]

[1,2,3,4, 5] Department of Computer Science and Engineering,
[1, 2,3,4,5]QIS College of Engineering and Technology, Ongole, Andhra Pradesh, India
[1]heena.sk@qiscet.edu.in, [2]jayasri.t@qiscet.edu.in, [3]ramya.v@qiscet.edu.in
[4]sheema.sk@qiscet.edu.in, [5]karuna.d@ qiscet.edu.in
Corresponding Author Mail: qispublications@qiscet.edu.in

Article Info
Page Number: 129 - 140
Publication Issue:
Vol 68 No. 1 (2019)

Article History
Article Received: *09 September 2019*
Revised: *16 October 2019*
Accepted: *21 November 2019*
Publication: *28 December 2019*

Abstract

Today, the majority of communication and exchange across all corporate sectors occurs via email. Spam, or undesired mass mail, is expanding dramatically in volume at the same time that emails are becoming the primary means of information transfer. Phishing emails can be used to spread pornographic content, solicit private information from recipients, or market products and services. In light of this, it is crucial to develop an entire system for identifying spam that is based on semantic text categorization and incorporates URL-based filtering and NLP.A model with excellent performance and efficiency has been sought for after studying several machine learning approaches.

Keywords— Online correspondence, pushy sales, phishing, intelligent retrieval, Some of the concepts used include add-on, random forest, naive Bayes, text ranking, URL ranking, and cyber security.

INTRODUCTION

Inside the twenty-first century and the age of globalization, emails account for the bulk of communication and exchange across all corporate sectors. In 2019, 246 billion emails were sent and received every day, and by 2021, that amount is anticipated to rise to 320 billion emails. Out of these, 117.7 billion are personal emails and 128.8 billion are corporate emails. Email communication is a very effective, formal, and efficient way to send information for all institutions, including people, businesses, and governments.The subject line and body of these emails very well may include highly sensitive information that is incredibly important to the company. That could include details about a potential, highly publicized transaction for a firm that cannot be exposed to the public, account and banking information for an individual, and details about the government that are not in the national interest.It cannot be emphasized how important end-to-end encryption is in this situation for all data transferred and received. The four main types of spam email that have been discussed in this article are aggressive marketing, pornographic emails, phishing, and email spoofing. Four components make up the document. first part contains the next section of the literature review describes the several machine learning methods that were employed, examines the algorithms, and applies the machine learning model to text categorization.

We developed URL filtering in the fourth stage, and in the fourth, we paired classification tasks with URL filtering to provide a Gmail add-on.

NATURE PRESERVE

Greater accuracy in [2] is provided by an integrated approach combining all three methods than by any technique utilizing a single strategy (URL Analysis, NLP, and ML). [3] More specifically emphasizes the URL Classification technique. Using a data set from the Phish tank, the decision tree method is utilized to train the model. [4] Evaluates the URL based on a variety of metrics, including the quantity of dots and unusual characters. The random tree constructs the model more rapidly than the KNN, making it the most useful machine learning approach for categorizing emails. The parameters for the random tree and the KNN are both equal in terms of numerical values. [5] The model was trained using data from the Enron Spam Project and makes use of the decision-making tree methodology, support vector machines, random forest, and K nearest neighbor methods.

An overview C add-on was also created in a visual studio. The most accurate machine learning algorithm was Random Forest. The model for text classification is first trained using data sets from the CMU Corpus and the Enron Email Corpus. In [6], file categorization is based on a total of 32 parameters. For something like the text classification and file classification stages, the Support Vector Machine method achieves the best results. [7] was familiar of the drawbacks of term frequency-based models, which have a heavy computational burden and sluggish training speed because of the size of the enormous feature vector space. Putting into practice a semantic similarity-based methodology that uses efficient information retrieval algorithms to peel back layers of semantic meanings. [8] Describes how several machine learning (ML) systems performed when identifying spam and fake internet reviews utilizing different components including keyword extraction, unigrams, bigrams, and n-grams.

The usefulness of different pre-processing strategies on classification accuracy was taken into account in [9], where the cosine similarity function measure was only used to specific parts of speech (POS). In [10], a semantic feature space was produced from training data by using statistical techniques, and this feature space also was utilised to implement the BP algorithm to address issues with traditional neural networks. A model for efficient feature selection that making use of a variable learning rate is applied to keyword-based spam filtering. Examines current software and the range of spam classification techniques, compares performance metrics of various supervised learning algorithms, and extracts email routing statistics from a spam source. [11] Researches current spam statistics and various spam attacks (email phishing, spear phishing, and spoofing).[12] focused more and more on SMS spam control using text mining technologies like Rapid Miner for identification and clustering algorithms. Because they perform effectively even in the lack of significant data and feature engineering and modelling, SVMs and Delusional Bayesian models have been justified. The prime objective of [13] is to test and contrast four different machine learning algorithms for content-based spam filtering methods.Using a neural network classifier alone to reject spam is not recommended, according to experiments, as it is more susceptible to the size of the training set. Compared to the Naive Bayesian classifier, the SVM and RVM classifiers perform significantly better in general and are less sensitive to feature sizes and data sets. The design and implementation of back-propagation neural networks for spam segmentation using behavior-based features are shown in [14]. Further, a rule-based method is

presented for instantiating behavior-based features into discrete values. The research also analyzes the features used to categorize spam and email spamming conduct. In [15], where they also emphasis the relevance of picking certain characteristics, the adjustment of two Random Forest parameters is done to enhance the spam detection rates.The RF-based spam classifier makes use of both feature selection and parameter optimization. It can detect spam with high accuracy and low computing power. Several cutting-edge tactics for various spam filtering approaches were looked at in the [16] study. Gmail, Yahoo, and Outlook all utilise different spam filtering algorithms; the basis for assessing performance is various performance assessment measures.The study contains a comparison of all machine learning algorithms used for filtering as well as a list of outstanding research issues that current methods must address. [24] Highlights a method that used SVM and KNN for the classification of Chinese web pages, improving the predictability of the classification and providing better feedback.

METHODOLOGY

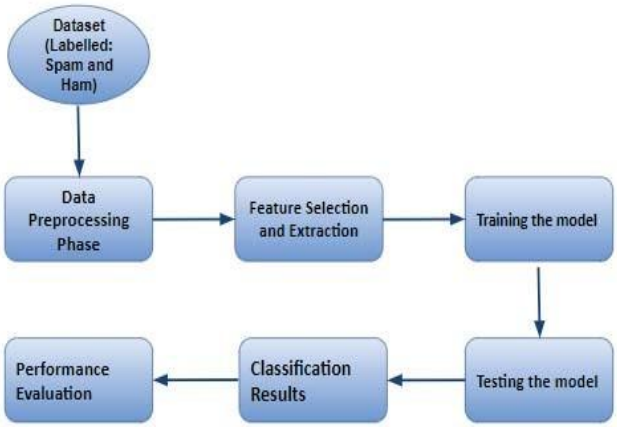

Figure1:FunctionalBlockDiagram

A. Figure 1 shows a block structure of the approach and procedures used to implement several machine learning algorithms for text categorization.

Implementedwhichareasfollows:

- Maximum likelihood estimate using Naive Bayes
- Euclidean distance: KNN
- • Number of nodes and class weight in a decision tree
- Maximum features Random Forest

• Decision function shape a support vector machine

1) Using Bayes' theorem for conditional probability and two presumptions, Naive Bayes is a classification technique that employs probabilistic classifiers.all characteristics are equally significant and all features are independent of one another. When modelling the features for text classification, a multinomial distribution or, if the characteristics are continuous, a Gaussian distribution, may be used.A Nave Bayes classifier is often used for classifying text in the feature vector space because it performs well with high-dimensional data points [8], learns more quickly, requires less training samples, and does so. Its performance is on par with that of more complex approaches with the proper data pre-processing.

1) K-Nearest Neighbors (KNN): This well-liked supervised learning method is primarily utilised for categorization. It assumes connected elements are placed near to one another. Using essential

inputs the Euclidean Length degree of similarity is evaluated. Since a model doesn't need to be established or its parameters adjusted, KNN is simple to implement. KNN does not assume anything about the distribution of the data because it is a non-parametric procedure. However, as the dataset's dimensions and size develop, the process executes more slowly.

Decision Tree: Decision trees are popular machine learning tools that are used for both classification and regression. Their tree structure resembles a flowchart. It makes use of recursive partitioning.

A technique for identifying characteristics or attributes using certain purity indices [8]. Entropy and gain are the indices that are most frequently employed. Entropy, or degree of uncertainty, is measured by the Gini index, which assesses the probability that a feature chosen at random would be misclassified. Relates to the increase in information. These are employed to choose the attributes that should be located at the root, internal node, or internal node. In decision trees, categorical and continuous variables can both be employed. When adapted for our application, the decision tree approach is shown in Figure 2 in action.

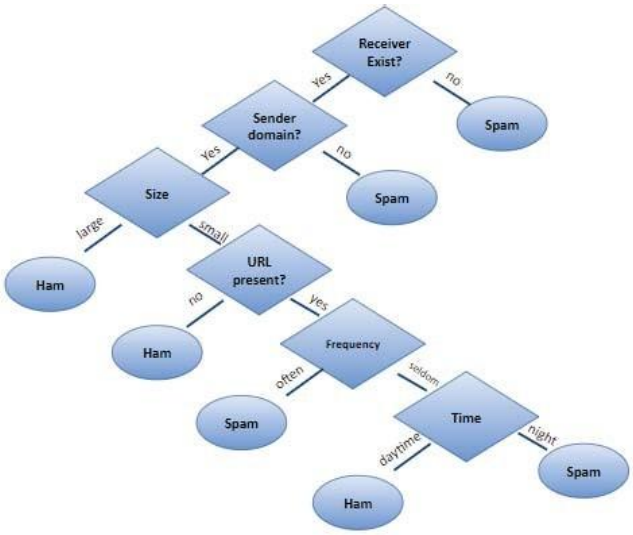

Figure2:DecisionTreeFlowchart

1) Randomized Forest: Random Forest is an ensemble learning technique that uses a sizable percentage of uncorrelated decision trees to build a model that performs well on fresh datasets because of its increased generalization to surprising, novel data.By employing the Bootstrap Aggregation or Bagging strategy, which divides the training set's data into many subsets and selects them at random with replacement [19], it lowers variance and always guards against over-fitting the model. Furthermore, every decision tree is trained using these subgroups. A good performance on fresh datasets is obtained from predictions made by a large number of uncorrelated models. Figure 3 shows a collection of decision trees that represent the random forest approach.

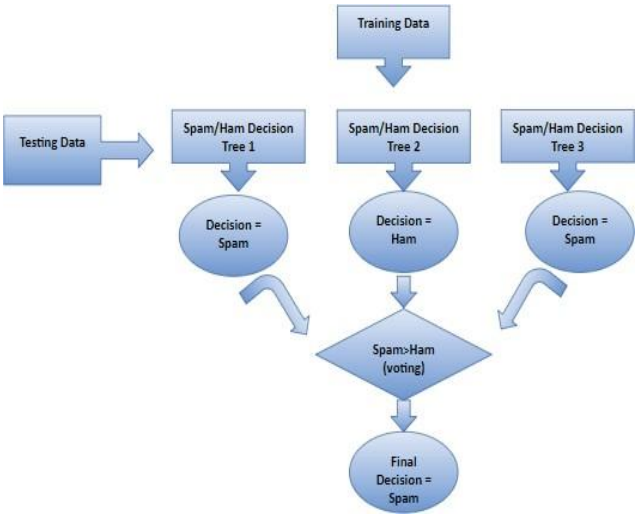

Figure3:RandomForestFlowchart

B. Support vector machines (SVM) are some of the most used supervised methods for segmentation. Based on the model's predictions, an n-dimensional workspace is used to create a hyper plane that separates each class from the others by examining the qualities in the training data set [19]. The best decision surface is produced using SVM, a classifier with a large margin, in order to have the greatest separation from the training sets for each class. SVM is distinctive in that it can convert non-linearly separable data into linearly separable data utilizing kernel methods.Kernel techniques allow for operation in higher dimensions by computing the inner products of the vectors rather than their actual coordinates. Kernel-based methods are used to transition training data from a non-linear decision surface into a linear decision boundary in a higher dimensional space. Support Vector Machines are economical computationally because they can handle bigger training sets. Text classification, sometimes referred to as text tagging and text categorization, is a method that divides texts into categories that may be unstructured or organised, depending on the needs. Numerous machine learning models employ natural language processing to analyse text, perform additional operations, and then categorize the results based on content. The flowchart and techniques forcategorising text are depicted in Figure 4.

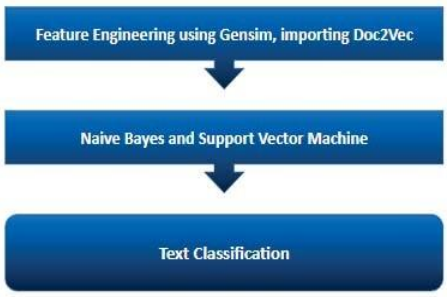

Figure4:TextClassificationflowchart

- Another popular open-source NLP tool for subject modelling is called Generate Similar, sometimes known as Genism. It is employed for a variety of challenging tasks, including handling sizable text collections, the semantic similarity approach, and the construction of word embedding

using Word2Vec to create feature vector spaces with 100 dimensions [19].

- This same data can go through a variety of pre-processing and cleaning steps in pipelines for natural language processing. The following is a list of them: Reducing tags: The text data collection may contain components that are not being utilized or are of no use, such as HTML. These tags are eliminated as a result.

• Eliminating accented characters: The collection of text data may contain accented characters like e', a', and n, among others. English is the primary language of the data set. Since this is the case, certain characters are converted to their standard ASCII values; for example, the character "a" becomes "a."

• Growing muscle spasms: The abbreviated words and syllables in the text data set are altered to their extended, original, and conventional forms, such as "I'd" is changed to "I would," "don't" is changed to "do not," etc.

Special non-alphanumeric characters might be identified in the text data collection, thus they should be removed.

for instance, $, @, etc., are pointless. They are removed using regular expression (regex).

Lemmatization and word stemming: By adding suffixes or prefixes to the words, word stems, which are the fundamental forms of words, are created. For example, the words dance, dancer, and dancing are all variations of the word dance. The method of learning the fundamental word is known as stemming. Lemmatization, like stemming, requires deleting affixes in order to get the root word rather than the root stem.

- Removing stop words: When generating features, many of the keywords in the data set have little to no meaning. Examples of these terms that tend to be used most frequently are a, an, the, is if, etc. A stop word list is one of the tools in a toolbox for natural language processing.

- Getting rid of stop words: Many of the keywords in the data set are either insignificant or useless for feature building. A, an, the, is if, etc. are the words that emerge most rarely. One of the tools in a set used for natural language processing is a stop word list.

Removing stop words: When creating features, many of the keywords in the data set have little to no meaning. Examples of these terms that tend to be used most frequently are a, an, the, is if, etc. One of the tools in a set for natural language processing is a stop word list.

Getting rid of stop words: Many of the keywords in the data set are either insignificant or useless for feature creation. A, an, the, is if, etc. are the words that appear most frequently. One of the tools in a set used for natural language processing is a stop word list.

Support vector machines: To produce results with incredibly high precision, support vector machines use an optimization technique. It classifies by identifying the most advantageous hyper plane and widening the gap between spam and ham. Due to SVM's robustness, it performs remarkably well when there are more dimensions than samples [16].

Victimization: We use vector space models, or VSMs, to overcome the aforementioned problems. Based on contextual and semantic similarity, these representations are used to show word vectors in n-dimensional vector space. The semantic similarity of words that exist and are used in the same context is well established. We make use of word2vec models to do this.

Clear and simple, word2vec is a natural language processing approach that first learns word associations from a dataset using a two-layered neural network model. Then it can predict sounds meaning or recommend words to complete the statement. Every word in the data set has a specific

matching vector assigned to it, and these vectors are arranged so that words with comparable meanings or situations are put close collectively [19]. The Word2vec approach generates a vector space with 100 or more dimensions by ingesting a substantial quantity of text input. The model architecture predicts the current word for the aforementioned procedure.

Continuous bag of words: Using the exact words nearby or in the direct vicinity.

Continual skip-grams: These anticipate the words that will appear just after the solutions based or the words that will come after it in the phrase. Graph [20]

As stated previously, the text classification process is the initial step in the categorization of spam, and it leverages the most extremely accurate machine learning algorithms, Naive Bayes and Support Vector Machine, to classify text containing trigger words. Categorizing and filtering URLs are also steps in the subsequent phase.

To preserve data privacy, safety, integrity, and security, an important that staff understand must be utilized to safeguard clients from emails that include risky URLs in the email body. The graph below further divides this period into four distinct stages:

Fig5:Schematic for URL filtering

•URL Un-Shortening

•Similarity to URL on Blacklist

•Trigger Words Existing

•Special Characters' Identification

URL un-shortening: In rare cases, a spammer may consciously attempt to cloak the true harmful or malicious URL behind a URL that seems secure in an effort to trick the victim into supplying data or information. Shortening the original connection will result in a new link that seems safe from the outside, which the hacker may use to carry out this action. In order to acquire the original link and enable unidirectional redirection of people to it, the URL must first be un-shortened in order to overcome this problem. In the subsequent phases, this connection is evaluated,

Access to unwanted URL: After the original shortened URL and the unsharpened URL are compared to a list of URLs in the blacklisted URL data-set, the URL is officially considered malicious, and the email is classified as spam.

The URL is also analyzed for the presence of certain terms known as "trigger words." Such terms are listed in the spam trigger data-set [16], and if either phrase or word is contained, the URL and the email will be categorized as malicious and spam, respectively.

These phrases represent for a variety of mail, including phishing, marketing and advertising, financial scams, and pornographic terms.The very last stage in the process is the identification of any special characters that may be included in the URL. Attackers generally utilize these characters to generate URLs that resemble well-known, safe URLs, hoping that visitors would

mistake them for authentic sites and provide their personal information. In order to gain the user's trust, these URLs often link website traffic with a user interface resembling one that is well-known and trustworthy. This situation happens regularly for users, social networking sites, and the banking and financial sectors. Over this, it is possible to tell if a URL is malicious or not by checking at special characters like "@".

Furthermore, the URL will be included in the message and will be classified as malicious in this phase if any of the four sub-steps yields a positive result, i.e., if any feature is present.

AMALGAMATION

Following the individual development of models for the first stage, language processing, and the second, URL analysis and filtering, the next and penultimate stage included merging the two models to create a model that took into account the results of both the first and higher stage. Therefore, this case, we'll apply the OR operation.

The model was created to recognize an email as spam if any of the two outputs indicates that the text includes spam terms or if the URL is risky. A two-step procedure of text classification and URL classification must be performed before an email can be labelled as ham. After that, this model is made accessible via an API on the cloud server hookup platform.

The framework for creating software written in Google Apps Script is used to create user interfaces that may be connected to Google Workspace. By using the programming language for java script-based apps, developers may create a wide range of add-ons.

It is straightforward to test, deploy, publish, and connect add-ons with other Google products like Google Docs, Google Sheets, and Gmail thanks to the infrastructure that Google has given for developers.

An add-on for Google Workspace that connects with Gmail has been created as a demonstration. The programmer's software categorizes the emails as spam or ham in real time by sending a request to the API.RESULTS AND ANALYSIS OF EXPERIMENTS

Even though foremen slimmed down using the Term Frequency Inverse Document Frequency approach, text classification had been tried to carry outKey performance parameters including Accuracy, Precision, Recall, and F1 score were taken into consideration while comparing the performance of 5 different machine learning algorithms. To help with this, a confusion matrix has been developed. The implemented algorithms are listed below:

The algorithms are applied to two separate data sets: the spam.csv published on Cagle [23] and the spam data set from Enron [22]. Data set from Cagle: This data set's spam collection consists of a number of SMS messages with tags that we utilized for experimentation. This data collection includes 5574 English mails that have been categorized or labelled as legitimate or spam [23].Enron Data Set: In order to conduct experiments utilizing various ML algorithms and obtain the desired results, the Enron Data Set was employed. There are 30207 samples in the Enron data set overall, 16545 of which were identified as authentic or ham, while the remaining 13662 occurrences were labelled as spam [22].

The following 5 algorithms were added to a machine learning model and used for classification after being imported from the Sickie-Learn package. The following comparative tabular analysis is included in the table as a note:

Algorithm		Naive Bayes	Support Vector Machine	Decision Tree	Random Forest	K Nearest Neighbour			
	0/1					K=1	K=3	K=6	K=10
Precision	0	95%	98%	91%	96%	95%	93%	90%	89%
	1	97%	97%	83%	99%	99%	100%	100%	100%
Recall	0	100%	100%	99%	100%	100%	100%	100%	100%
	1	68%	86%	41%	74%	66%	50%	30%	24%
F1 Score	0	97%	99%	95%	98%	97%	96%	95%	94%
	1	80%	92%	55%	85%	79%	67%	46%	38
Accuracy	Model	95.48%	97.83%	90.90%	96.43%	95.29%	93.25%	90.58%	89.69

Table1:Tabular Comparison Analysis 1

		Naive Bayes	Support Vector Machine	Decision Tree	Random Forest
Precision	0	94%	96%	96%	98%
	1	93%	73%	91%	97%
Recall	0	97%	91%	97%	99%
	1	85%	86%	89%	95%
F1 Score	0	96%	93%	96%	98%
	1	89%	79%	90%	96%
Accuracy	Model	93.65%	89.55%	94.70%	97.60%

Table2:Tableau 2

		Naive Bayes (Multinomial)	Support Vector Machine	Naive Bayes (GNB)
Precision	0	92%	96%	90%
	1	96%	73%	92%
Recall	0	82%	91%	78%
	1	71%	86%	71%
F1 Score	0	84%	93%	76%
	1	82%	79%	84%
Accuracy	Model	95.8%	89.55%	89.8%

Table3:

Additionally, the Genism Library is used to fix several flaws in the TF IDF method. In this study, the TF-IDF technique was abandoned in favor of the genism package, which supports semantic similarity and produces superior results. In addition, a three step filtering and analysis process was used to classify the URL in addition to the text that was there.

\

FUTURE STATE

Multiple sub-domains can be used for more study on this subject. At first, employing a more computationally costly but precise machine, the emphasis might be on increasing accuracy. Learning classifiers as XG Boost. It is also possible to investigate various word embedding techniques besides Genism word2Vec. Deep learning research could make use of the 2017-introduced transformer-based deep learning models. It features pre-trained algorithms for text summarization and translation as well as the ability to train on enormous data sets. Last but not least, the present data sets do not priorities real-time learning of email classifiers. Real-time elements have a significant influence in determining the classification accuracy, hence it is significant.

CONCLUSION

A thorough and effective technique for classifying spam has been developed, and it uses a two-step process to verify whether the mail being received is spam or not. To identify whether any links contained in the email are malicious or not, text classification is done first, then URL analysis and filtering. Five machine learning methods were examined and tested for text classification, and the two with the best accuracy. The final model included the use of Support Vector Machine and Naive Bayes. Lists of prohibited URLs and spam trigger phrases have been compiled from a variety of data sources. This model was made available as an API to instantly classify emails in mail.

REFERENCES

[1] Statist, accessed 3 November2020,https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/

[2] E.Marková,T.BatesP. Skoltand"Classificationofmaliciousemails",2019IEEE15thInternationalScientificConference on Informatics, Poprad, Slovakia, 2019, pp. 000279-000284, doe:10.1109/Informatics47936.2019.9119329.

[3] M. S. Swathe and G. Sara, "Spam Email and Malware Eliminationemploying various Classification Techniques", 2019 4th InternationalConference on Recent Trends on Electronics, Information, Communicate Technology (RTEICT), Bangalore, India, 2019, pp. 140-145, do10.1109/RTEICT46194.2019.9016964.

[4] S.NandhiniandD.J.Marseline.K.S,"PerformanceEvaluationofMachine Learning Algorithms for Email Spam Detection", 2020 Inter-nationalConferenceonEmergingTrendsinInformationTechnologyandEngineering (ic-ETITE), Vellore, India, 2020, pp. 1-4, doi: 10.1109/ic-ETITE47903.2020.312.

[5] K. Kandasamy and P. Koroth, "An integrated approach to spam clas-sification on Twitter using URL analysis, natural language processingand machine learning techniques", 2014 IEEE Students' Conference onElectrical, Electronics and Computer Science, Bhopal, 2014, pp. 1-5,doi:10.1109/SCEECS.2014.6804508.

[6] S.B.RathodandT.M.Pattewar,"Acomparativeperformanceevaluationof content based spam and malicious URL detection in E-mail", 2015IEEEInternationalConferenceonComputerGraphics,VisionandInformationSecurity(CGVIS),Bhubaneswar,2015,pp.49-54,doi:10.1109/CGVIS.2015.7449891.

[7] WeiHu,JinglongDu,andYongkangXing,"SpamFilteringbySemantics-based Text

Classification", 8th International Conference onAdvanced Computational Intelligence Chiang Mai, Thailand; February14-16,2016

[8] Crawford, M., Khoshgoftaar, T.M., Prusa, J.D. et al. ,"Survey of reviewspam detection using machine learning techniques", Journal of Big Data2,23(2015).https://doi.org/10.1186/s40537-015-0029-9

[9] VladSandulescu, Martin Ester "Detecting Singleton Review SpammersUsing Semantic Similarity", WWW '15 Companion Proceedings of the24th International Conference on World Wide Web, 2015, p.971-97610.1145/2740908.2742570

[10] Cheng Hua Li, Jimmy Xiangji Huang "Spam filtering using semanticsimilarityapproachandadaptiveBPNN",NeurocomputingJournal, Elsevier,https://doi.org/10.1016/j.neucom.2011.09.036

[11] Krishnan Kannoorpatti, AsifKarim , Sami Azam, Bharanidharan Sanmugam,"onAComprehensiveSurveyforIntelligentSpamEmailDetection,"IEEEJournalofComputationalIntelligence,2015.

[12] Zainal K, Sulaiman NF, Jali MZ, "An Analysis of Various AlgorithmsFor Text Spam Classification and Clustering Using RapidMinerandWeka", (IJCSIS) International Journal of Computer Science and In-formationSecurity,Vol.13,No.3,March2015

[13] B.Yu,Z.Xu,"A comparative study for content-based dynamicspamclassification",Knowl.BasedSyst.,China,2008,doi:10.1016/j.knosys.2008.01.001

[14] C.H.Wu,"Behaviorbasedspamdetectionusingahybridmethodofrulebasedtechniquesandneuralnetworks",ExpertSystemswithApplications,Kaohsiung,Taiwan,2009,doi:10.1016/j.eswa.2008.03.002

[15] S.M.Lee, D.S.Kim, J.H.Kim, J.S.Park, "Spam Detection Using FeatureSelectionandParameterOptimization",2010InternationalConferenceonComplex,IntelligentandSoftwareIntensiveSystems,DOI10.1109/CISIS.2010.116

[16] E.G.Dada,J.S.Bassi,H.Chiroma,S.M.Abdulhamid,A.O.Adetunmbi,O.E.Ajibuwa,"Machinelearningforemailspamfiltering: re-view,approachesandopenresearchproblems",Heliyon(2019) DOI:doi.org/10.1016/j.heliyon.2019.e01802

[17] 455SpamTriggerWordstoAvoidin2019,accessed3November2020,https://prospect.io/blog/455-email-spam-trigger-words-avoid-2018/

[18] PhishTank,accessed3November2020,https://www.phishtank.com/

[19] Word2vecskipgramandcbow,accessed3November2020,https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

[20] AsifKarim, Sami Azam, BharanidharanShanmugam, Krishnan Kan-noorpatti, and MamounAlazab - "A Comprehensive Survey for Intelli-gent Spam Email Detection", College of Engineering, IT and Environ-ment,CharlesDarwinUniversity,Casuarina,NT0810,Australia.

[21] TwoSimpleAdaptationsofWord2VecforSyntaxProblems-ScientificFigureonResearchGate,accessed3November2020,https://www.researchgate.net/figure/Illustration-of-the-Skip-gram-and-Continuous-Bag-of-Word-CBOW-modelsfig1281812760

[22] EnronSpamdatasetaccessedon3November2020,http://nlp.cs.aueb.gr/softwareanddatasets/Enron-Spam/index.html

[23] Kaggledatasetaccessedon3November2020,https://www.kaggle.com/uciml/sms-spam-collection-dataset

[24] Y. Lin and J. Wang, ”Research on text classification based on SVM-KNN,” 2014 IEEE 5th International Conference on Software Engineer-ing and Service Science, Beijing, 2014, pp. 842-844, doi: 10.1109/IC-SESS.2014.6933697.
[25] Detecting Spammers on Twitter": Fabr´ cioBenevenuto, GabrielMagno, Tiago, Rodrigues, and Virglio Almeida. In Anti-Abuse and SpamConference (CEAS) (July 2010).
[26] P Ramprakash, M Sakthivadivel, N Krishnaraj, J Ramprasath. ”Host-based Intrusion Detection System using Sequence of System Calls” International Journal of Engineering and Management Research, Vandana Publications, Volume 4, Issue 2, 241-247, 2014
[27] N Krishnaraj, S Smys.”A multihoming ACO-MDV routing for maximum power efficiency in an IoT environment” Wireless Personal Communications 109 (1), 243-256, 2019.
[28] N Krishnaraj, R Bhuvanesh Kumar, D Rajeshwar, T Sanjay Kumar, Implementation of energy aware modified distance vector routing protocol for energy efficiency in wireless sensor networks, 2020 International Conference on Inventive Computation Technologies (ICICT),201-204
[29] Ibrahim, S. Jafar Ali, and M. Thangamani. "Enhanced singular value decomposition for prediction of drugs and diseases with hepatocellular carcinoma based on multi-source bat algorithm based random walk." Measurement 141 (2019): 176-183. https://doi.org/10.1016/j.measurement.2019.02.056
[30] Ibrahim, Jafar Ali S., S. Rajasekar, Varsha, M. Karunakaran, K. Kasirajan, Kalyan NS Chakravarthy, V. Kumar, and K. J. Kaur. "Recent advances in performance and effect of Zr doping with ZnO thin film sensor in ammonia vapour sensing." GLOBAL NEST JOURNAL 23, no. 4 (2021): 526-531. https://doi.org/10.30955/gnj.004020 , https://journal.gnest.org/publication/gnest_04020
[31] N.S. KalyanChakravarthy, B. Karthikeyan, K. Alhaf Malik, D.BujjiBabbu,. K. NithyaS.Jafar Ali Ibrahim , Survey of Cooperative Routing Algorithms in Wireless Sensor Networks, Journal of Annals of the Romanian Society for Cell Biology ,5316-5320, 2021
[32] Rajmohan, G, Chinnappan, CV, John William, AD, ChandrakrishanBalakrishnan, S, AnandMuthu, B, Manogaran, G. Revamping land coverage analysis using aerial satellite image mapping. Trans Emerging Tel Tech. 2021; 32:e3927. https://doi.org/10.1002/ett.3927
[33] Vignesh, C.C., Sivaparthipan, C.B., Daniel, J.A. et al. Adjacent Node based Energetic Association Factor Routing Protocol in Wireless Sensor Networks. Wireless PersCommun 119, 3255–3270 (2021). https://doi.org/10.1007/s11277-021-08397-0.
[34] 9. C ChandruVignesh, S Karthik, Predicting the position of adjacent nodes with QoS in mobile ad hoc networks, Journal of Multimedia Tools and Applications, Springer US,Vol 79, 8445-8457,2020