ISSN: 2581-3404 (Online) **IF: 5.445 (SJIF)**

IJIRTM, Volume-7, Issue-5, October-2023.

An Email Spam Detection and Classification Framework Based On Machine Learning Algorithm

Nitish Kumar[1], Chetan Agrawal[2], Darshna Rai[3]

Department of Computer Science and Engineering, RITS, Bhopal, INDIA[1, 2, 3]

nitishkumarbe@gmail.com[1], chetan.agrawal12@gmail.com[2], darshna.rai@gmail.com[3]

Abstract: *In modern times, email has emerged as the most common way of communication that is also the quickest and most cost-effective. For many inhabitants, the data sharing process has evolved into a routine part of their day-to-day lives. Email, on account of its user-friendliness, is vulnerable to a significant deal of potential threats. The sending of unsolicited commercial email, also known as spam, is one of the most fundamental risks associated with electronic communication. The expansion of the spam movement is an issue that should be of concern since it wastes system data transfer capacity, the time of customers, and memory space, and it also results in monetary loss for both customers and businesses. In this research, we make a recommendation for a model that does characteristic choice for ruthless spam recognition in email. Our goal is to streamline the grouping parameter, the figure precision, and the computation time for later on characterization calculations. In order to accomplish this, we offer a model that can perform characteristic choice. The method of determining features will make use of the work of ling spam dataset, and after that, the arrangement of chosen features will be validated with four different classifiers, namely Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression, and Random Forest.*

Keywords: Email – spam, Machine learning, Spam Detection Classification.

Introduction

At present, electronic mail has emerged as one of the most crucial modes of communication. Regrettably, with the growing importance of messaging, there is a concurrent increase in the volume of unsolicited messages that are transmitted to users. Spam emails are unsolicited messages that are disseminated to a large number of recipients. The atypical functionalities of unsolicited emails are noteworthy. Some of these platforms address promotional concerns, while others priorities safeguarding against malware and phishing attempts that aim to compromise users' financial information.[1]

Recently, unsolicited commercial or spam email has emerged as a significant challenge on the internet. The act of spamming is deemed to be a futile exercise that consumes valuable storage space and transmission bandwidth for correspondence data. The phenomenon of unsolicited commercial email, commonly known as spam, has been a growing concern for a significant period. According to recent findings, approximately 40% of all messages are classified as spam, resulting in an estimated 15.4 billion emails being sent daily, and incurring a cost of approximately $355 million annually for internet users. Currently, programmed email filtering appears to be the most effective approach in combating spam, and there exists a fierce competition between spammers and spam-filtering methods. In the past, a common approach to managing

spam was to block messages from specific locations or filter out messages with particular subject lines. The utilization of certain dubious techniques by spammers has been observed in order to overcome filtering strategies. These techniques include the use of random sender addresses and the attachment of arbitrary characters to the beginning or end of the message subject line [2, 4].

Machine learning and information aggregation are two widely used methodologies in email filtering. When studying design methodology, it is necessary to establish a set of guidelines for categorizing messages as either spam or ham. It is recommended that the establishment of such standards be delegated to either the filter user or another qualified professional, such as the software company that provides the models. This is due to the fact that the standards require constant updates and maintenance, which can be a time-consuming and impractical task for most users. The effectiveness of the machine learning approach surpasses that of the information building approach as it does not necessitate the specification of any principles [4, 5]. The aforementioned set of practice assessments pertains to a series of pre-arranged electronic mail correspondences. Subsequently, a specific computation is employed to incorporate the classification regulations derived from the aforementioned electronic mail correspondences. The utilization of various algorithms in email filtering has been extensively explored within the context of machine learning methodology.

This study utilizes an individual's email record containing a quantity of messages exceeding the typical amount for the purpose of analysis and assessment. The main objective is to comprehend the email content and address, and categories all emails into two distinct groups: private and professional, based on factors such as content, sender, and subject line. The determination of whether messages are classified as spam or ham requires the implementation of specific methodologies and strategic approaches.

This rest of this paper is prepared as follows. Section 2 describes the literature review of various machine learning methods used in classifications. Section 3 describes proposed work in which we explain algorithms for e-mail classification and its flow chart, and Section 4 presents the results of the analysis and discusses the relative algorithms for e-mail classification along with we represent them using confusion matrix as well as its accuracy. Section 5 concludes the paper by discussing the implications of the results for e- mail classification and analysis, and also explains future scope of this work.

II. Literature Survey

The dependence of individuals on social media in their daily routines is progressively escalating. The expansion of these platforms has been accompanied by a corresponding increase in our communication through them. Twitter is a highly utilized social media platform in the Middle Eastern region. Similar to other social networking platforms, Twitter is vulnerable to the presence of spam accounts that disseminate a portion of their tactics. Arab countries have experienced a significant dearth of suitable technological resources catering to the Arabic language, rendering them particularly vulnerable to cyber-attacks. In addition, the complexity of Arabic as a language with diverse dialects and intricate grammatical structures may pose challenges in the retrieval of textual data. Numerous recent studies have investigated innovative methods to mitigate spamming in tweets. The study gathered Arabic datasets that are suitable for identifying junk in order to address the issue of detecting spamming identities on Twitter in the Arab region. By employing Twitter's subscription API, the database incorporated data derived from premium content. Unclaimed personas were identified and data was categorized. A novel methodology utilizing deep learning algorithms has been devised, which offers several advantages, including enhanced efficiency and expeditious outcomes, while simultaneously minimizing system resource utilization. The study utilized a convolution neural network (CNN) methodology for conducting text-based statistical analyses, and a basic neural network-based framework for information analysis. Upon combining the results of both algorithms,

identities were classified as either spam or non-spam. The proposed methodology outperformed the most optimal designs assessed in the study, attaining a 94.28% efficiency rate through the utilization of a hybrid algorithm that leveraged superior extracted features. On social media platforms such as Twitter, a convergence of various Arabic accents and informal idiomatic expressions can be observed in online interactions. The increased complexity in identifying spam identities based solely on textual characteristics necessitates the implementation of multiple preliminary procedures to ensure accurate categorization. Further investigation into a preprocessing phase that could effectively handle Arabic accents with minimal impact on meaning and comprehension would have been advantageous. The primary constraint of spam accounts in conjunction with the text and meta-based deep learning framework suggested by [9] is deemed significant.

The categorization and recommendation approach for identifying the preferences of social networking site (SNS) members is a crucial aspect in various industries, particularly in marketing. Tailored advertisements aid companies in standing out amidst the vast expanse of digital marketing by enhancing their relevance to consumers and evoking favorable responses [10]. The thorough assessment of images and messages in user posts has the potential to provide more precise predictions of user preferences. It is noteworthy that previous studies on user preference classification have predominantly focused on textual data. Consequently, the present investigation employs a combination of linguistic and visual elements to categories the preferences of social networking site users. In a recent study, researchers evaluated various convolutional neural network (CNN) and recurrent neural network (RNN) models for user preference categorization systems. The Curlie directory was utilized to identify consumer interests. [11] The categorization of photos through individuals' social networking service (SNS) posts has been achieved through the application of convolutional neural classification methods. On the other hand, the classification of textual information has been accomplished through the utilization of RNN-based classification methods in their hybrid neural network system. The study's results indicate that the most effective method for categorizing users' preferences was through the combination of text and graphics, which achieved a success rate of 96.5%. In comparison, using text alone resulted in a success rate of 41.39%, while using photos alone resulted in a success rate of 93.2%. These findings were obtained through rigorous testing. The proposed methodology facilitates marketers in generating interest-based, ranked-order, and real-time suggestions by providing insights into customized social networking service (SNS) marketing communications. According to their interpretation, this article represents one of the initial attempts to utilize a combination of image and message statistics derived from user-generated content in order to improve the accuracy of identifying the political preferences of social networking site users. The ultimate goal of this approach is to enhance the efficacy of targeted advertising efforts. [12]

In recent times, single-modal spam filtering algorithms have achieved optimal classification performance for both image and text-based unsolicited messages. In order to evade detection by spam filters that rely on single-mode analysis, scammers incorporate extraneous information into the multi-modal component of an email and integrate it in a manner that reduces the classification performance of single-modal spam detection processor architectures. This enables them to successfully evade identification. The latest product, referred to as multimodal design, has been introduced as a solution to efficiently filter trash that may have been concealed in word or phrase. This product utilizes a multimodal fusing strategy and is obtained from numerical fusing (MMA-MF). The methodology employed involves the integration of a convolutional neural network (CNN) with a long short-term memory (LSTM) framework for the purpose of filtering waste. The e-mail's visual and textual elements were separately analyzed using long short-term and convolutional neural network models to produce two posterior distribution categorizations. These categorizations were subsequently combined into a hybrid framework to determine whether the message is

classified as spam or not. The utilization of a grid search optimization algorithm is being employed by researchers to ascertain the optimal hyper-parameters for the MMA-MF designer's hyper-parameters. Additionally, a k-fold cross-validation technique is being utilized to evaluate the algorithm's efficacy. The experimental results indicate that the aforementioned approach exhibits superior performance compared to conventional spam detection algorithms, achieving accuracy rates ranging from 92.65% to 98.49%. The researchers propose utilizing a new methodology, along with a one-class classification algorithm and few-shot pedagogical practices, to tackle the issue of the disparity between spam and non-spam emails. Additionally, they plan to gather additional realistic combined electronic mail data sources to improve the system framework of the prototype, thereby enhancing its effectiveness in detecting phishing attempts [13].

The proliferation of unsolicited bulk e-mail, commonly known as spam, poses a significant challenge to e-mail account users in accessing essential information. Various methods have been employed to implement email spam prevention in public mail servers. Nevertheless, it should be noted that certain email platforms do not provide the option to filter out unsolicited emails that contain valuable information, particularly for a restricted number of corporate email accounts. In order to protect email users from unsolicited messages, it is recommended that the system administrator establish a separate or integrated mechanism for detecting and filtering junk mail. The objective of this study is to determine the optimal approach for detecting unsolicited and unwanted electronic messages commonly referred to as spam. The researchers employed machine learning methods, including decision tree, logistic regression, and random forest, to ascertain the optimal approach for detecting spam emails. The results were subsequently assessed. The efficacy of spam message detection is evaluated based on the swiftness of train and test protocols, as well as the dependability of the outcomes. Based on the results of this study, it can be concluded that the implementation of the random forest methodology yielded exceptional results, as evidenced by a testing data processing time of 0.19 seconds and a reliability rate of 98 percent. The aforementioned discovery has the potential to serve as a foundation for the development of diverse spam filtering algorithms. The research is limited by the researcher's assumption that the utilization of more specific algorithms, such as the approximate solution and the database process, would lead to an improvement in efficiency [14].

Smart objects are the primary providers of computational services in close proximity to end-users for the upcoming Internet of Things. These devices equipped with embedded intelligence have the capability to make independent decisions within their respective environments by utilizing various artificial intelligence techniques. In light of the aforementioned challenges, scholars suggest a cognitive incursion prevention mechanism aimed at thwarting the infiltration of brand loyalty into the image data of web address bars. This mechanism is crucial in preserving the authenticity of search engine result pages. The proposed model offers ambient knowledge for web data filtering and detection of web spam by targeting three distinct levels, namely, data collecting services, edge computing services, and cloud services. The objective is to detect images that may cause harm. The proposed framework involves retrieving the average, image gradient, and volatility of a picture, followed by subsequent analysis of the obtained information. Deep learning techniques are being employed to evaluate the performance of the proposed method. The model demonstrated a 98.77% level of accuracy upon being evaluated using a dataset that was available in real-time [15].

Unsolicited emails, commonly referred to as spam, have the potential to result in significant losses for both email recipients and servers when utilized by spammers. Thus, the implementation of a sophisticated email spam classification system is essential for the purpose of identifying and preventing unsolicited emails from infiltrating our inbox. The present study introduces two commonly used machine learning techniques, namely NaiveBayes Classifier and Support Vector Machine, for the purpose of categorizing emails as either spam or ham, depending on the content or body of the emails. The Naive Bayes Classifier regards

individual words as independent features. The utilization of Support Vector Machine is viable in the representation of an email within a vector space, where each feature corresponds to a distinct dimension. The study concludes by conducting a comparative analysis of two distinct methods based on their performance metrics, namely precision, recall, and F-measure. The objective is to determine the optimal method [16].

III. Proposed Work

In this Research paper we have proposed a machine learning method for categorization computations that obtain by attempt at data analysis. We took ling spam corpus dataset for the purpose of experiment, which is incredibly enormous dataset and it includes various emails and these emails are categorized to organize emails and analyze the emails, which is enlighten through figure 1. Here now we initially describe regarding how ling spam corpus dataset preprocessed step by step.

3.1 Ling Spam corpus dataset

Initially categorization computation is compared based on confusion matrix and precision. These computations of categorization have been employed on the ling spam corpus dataset, which contains of enormous emails for the purpose of training and testing. The step by step process involved in this are as following:

- Initially organize the dataset by preprocessing of it
- Lexicon table will be formed for all word
- Do Feature extraction
- Train the classifier and then testing

3.1.1. Organize the dataset

During this procedure we have to preprocess the dataset. For this purpose we have acquired the ling corpus dataset which consist of 702 training emails and 260 testing emails, so we have collectively approximately 962 emails.

a) **Stop Words Removal –** Stop words like “the”, “and", "of", and so on are very common words in English sentences over and above they aren’t really significant for detecting ham or spam emails thus these words have been separate initially from the emails.

b) **Stemming –** during this process, compile mutually the different altered types of a word so they could be scrutinize as a meticulous article. For instance, "include", "includes," and "included" would all be symbolized as "include". The circumstance of the sentence is similarly conserved in lemmatization as contrasting to stemming (another term in text mining which doesn’t consider denotation of the sentence).

3.1.2. Lexicon table Formation

Generally the main line of the email is subject and the third line includes the main body of the email. We now carry out word analysis on the content to identify the spam emails. As primary step, we must produce a lexicon of words and their occurrence. For this work, training of 700 emails is utilized.

Once the lexicon is produced we could embrace only a small number of lines of code to the above ability to eliminate non words.

3.1.3. Feature extraction

Once the lexicon is organized; we could acquire word count vector of 3000 dimensions proposed for all email of training set. All word ensured that the vector hold the frequency of 3000 words in the training. Obviously we might have estimated at this point a vast segment of them determine be zero. For example,

suppose we have 500 words in our lexicon. All words verify the vector enfold the frequency of 500 lexicon words in the training file. Suppose text within training was "Get the work done, work done" then it will be pre-assembled as:
[0, 0, 0, 0, 0,… ….0,0,2,0,0,0,… … ,0,0,1,0,0,… 0,0,1,0,0,… … 2,0,0,0,0,0]
Here, each one of the word counts are placed at 296th, 359th, 415th, 495th catalog of 500 length word count vector additionally to the remaining will zero.

3.1.4. Classifiers Training

We trained the dataset by four machine learning algorithms i.e. Naive Bayes, Support Vector Machines, Logistic Regression and Random Forest.

- The Logistic Regression is renowned machines learning algorithm for twofold classification. It acquires actual weighted sources of data and constructs a prediction with consideration to the probability of the data residing to its default class.
- Naive Bayes is a conventional classifier with enormously renowned technique proposed for text mining problem. It is a supervised probabilistic classifier based on Bayes theorem cooperative between every match of characteristics.
- The decision trees night different by limiting the attributes that the greedy computation could appraise at each split instant that creating the tree. This is known as the Random Forest computation.
- SVMs are supervised paired classifiers which are incredibly efficient when you have superior number of attributes. The purpose of SVM is to split a few subset of training data from remaining dataset is known as the support vectors. The decision capability of SVM model that forecasts the group of the test data is based on support vectors and formulates utilization of a kernel.

Once the classifiers are trained, we could test the performance of the proposed models on test dataset. We eliminate word comprise vector matrix for all email test dataset with forecasting its group i.e. ham or else spam, through the various machine learning algorithms NB classifier, Logistic regression, SVM model, Random forest.

3.2. Proposed Algorithm

Test-set include 130 spam emails as well as 130 non-spam emails. If you have approach so far, you will discover below result. I have exposed the confusion matrix of the test-set intended for together the models. The diagonal parts represent the accurately known (true identification) mails where as non-diagonal element represents incorrect classification (false identification) of mails.

Algorithm:
Input: Ling Corpus Data set
Output: Result every of the Classifier
1. Download the dataset commencing website or else from inherent library.
2. Preprocess the data set by preprocessing method
3. Subsequent to preprocessing relate a variety of machine learning classification algorithm similar to NaiveBayes, logistic regression, random forest and SVM lying on the preprocess dataset.
4. Work out the accuracy of the classification technique.
5. Evaluate every one of the classification technique.

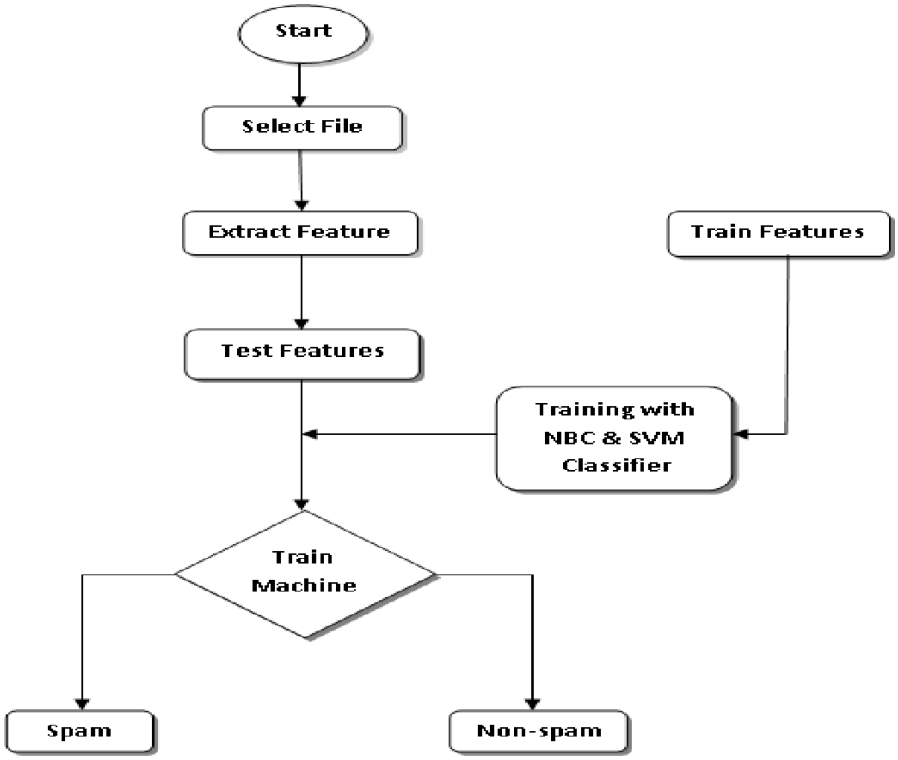

Figure: 1. Proposed method flow graph for classification via NBC & SVM.

IV. Result Analysis and its Parameters Metrics Used

Here, we use the Python version 3.6 for examination as well as its parameter which is use of this investigation. The series of steps and every one of the computations with it will be showed in this segment, in both parallel and sequential evaluation. The best four models for evaluations are likewise presented here.

4.1. Confusion Matrix:

A confusion matrix is an abstract of prediction outcome on a classification difficulty.

The numeral of correct along with wrong predictions are sum up with count values as well as broken down through every class. This is the key in to the confusion matrix. The confusion matrix demonstrates the method in which your classification model is confused while it makes predictions. It gives us approaching not only into the errors individual made through a classifier but further prominently the kind of errors that are being made.

	Set 1 Predicted	**Set 2 Predicted**
Set 1 Actual	TP	FN
Set 2 Actual	FP	TN

Here, Set 1: Positive
Set 2: Negative

Explanation of the Terms:

Positive (P): Examination is positive (for case: is an apple).
Negative (N): Examination is not positive (for case: is not an apple).
True Positive (TP): Examination is positive, along with is predicted to be positive.
False Negative (FN): Examination is positive, other than is predicted negative.
True Negative (TN): Examination is negative, along with is predicted to be negative.
False Positive (FP): Examination is negative, other than is predicted positive.

4.2. Classification Accuracy:

Classification Accuracy is known through the relation:

Though, there are problems through accuracy. It assumes equivalent costs for mutually type of errors. A99% accuracy may be excellent, good, middling, poor or else awful depending leading the problem.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here, we have use classification algorithms available within particular library. Initially, we will estimate the confusion matrix subsequent that we will calculate the accuracy during by function or confusion metrics. accuracy_score; firstly we will demonstrate the output of Ling Spam Dataset which is specified here:

According to Naive Bayes

	Ham	**Spam**
Ham	129	1
Spam	9	121

Accuracy Score: 96.1538461538%

According to SVM demonstrate the confusion matrix

	Ham	**Spam**
Ham	126	4
Spam	6	124

Accuracy Score: 96.1538461538%

According to logistic regression, we will illustrate the output of Ling Spam Dataset which is given below:

	Ham	**Spam**
Ham	126	4
Spam	1	129

Accuracy Score: 98.0769230769%

As indicated by Random forest

	Ham	**Spam**
Ham	124	6
Spam	6	124

Accuracy Score: 95.3846153846%

Support Vector Machine, Naïve Bayes, Logistic Regression and Random Forest classifier were implemented and compared to each other in terms of accuracy score. The comparison of classifiers results are shown in the following table.

Method	**Base Methods**	**Data Set (Ling-Spam Corpus)**
SVM	91 %	96.15 %
Naive Bayes	92 %	96.15 %
Logistic Regression	-	98.07 %

Table 1: Comparisons of previous and present result on given data set.

Comparative study of based methods to be used in previous paper and proposed methods in given table 1:

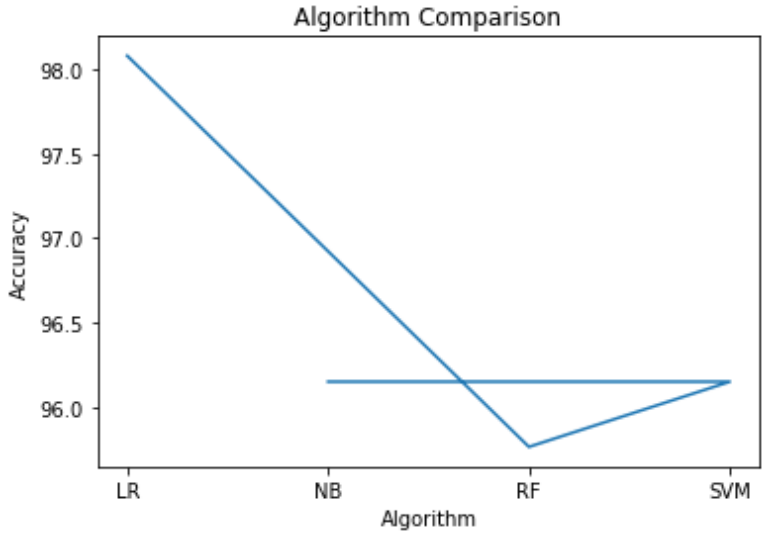

Figure: 2. A Comparative Study of different classifier on ling_spam_corpus Dataset.

Accurate classification results for every classification methods lying on ling_spam_corpus Dataset as well as comparisons to all other with accuracy are shown within Figure 2.

V. Conclusion

This paper provides a comprehensive overview of the prominent machine learning techniques and their applicability to the issue of spam email classification. The paper presents a detailed account of the calculations employed and their correlation with the Ling corpus Spam Dataset. The experimental results are highly encouraging, particularly with regard to the lesser-known calculations used in commercial e-mail filtering packages. The spam recall percentage in the five methods is accurately measured, and the NaiveBayes, SVM, and Logistic Regression methods demonstrate superior performance in terms of accuracy. Further research is required to improve the performance of the NaiveBayes classifier, either through a hybrid system or by addressing the issue of feature dependence. Alternatively, the Immune system could be hybridized through harsh sets. Currently, hybrid systems are considered the most effective method for developing a successful anti-spam filter. Future endeavors will focus on achieving precise classification, with a complete absence of misclassification of Ham E-mail as Spam and Spam E-mail as Ham. Efforts will be directed towards mitigating the issue of phishing attacks, specifically in relation to

phishing e-mails, which have become a growing concern in recent times. Furthermore, the scope of this study can be expanded to address the issue of preventing Denial of Service (DoS) attacks, which have recently manifested in a distributed form known as Distributed Denial of Service (DDoS) attacks.

References

[1] Issam dagher, Rima Antoun," Ham- Spam Filtering Using DIFFERENT PCA SCENARIOS", 2016 IEEE International Conference on Computational Science and Engineering, IEEE International Conference on Embedded and Ubiquitous Computing, and International Symposium on Distributed Computing and Applications to Business, Engineering and Science
[2] Sch¨olkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)
[3] Ali, S., Smith-Miles, K.A.: A meta-learning approach to automatic kernel selection for support vector machines. Neurocomputing 20(1-3), 173–186 (2006).
[4] Spam (electronic), http://en.wikipedia.org/wiki/Spam_%28electronic%29 Vapnik, V.: Statistical Learning Theory. John Wiley and Sons (1998).
[5] Li, K. and Zhong, Z., "Fast statistical spam filter by approximate classifications", In Proceedings of the Joint international Conference on Measurement and Modeling of Computer Systems. Saint Malo, France, 2006.
[6] D. Heckerman and M. P. Wellman, "Bayesian networks," no. 3, March 1995, pp. 27–30.
[7] S. Whittaker, V. Bellotti and P. Moody, "Introduction to this special issue on revisiting and reinventing e-mail", Human-Computer Interaction, 20(1), 1-9, 2005.
[8] E-mail spam, http://en.wikipedia.org/wiki/E-mail_spam
[9]. A. S. Rassam and M. A. Rassam, "A combined text-based and metadata-based deep-learning framework for the detection of spam accounts on the social media platform twitter," Processes, vol. 10, no. 3, p. 439, 2022.
[10]. A. I. Taloba, R. Alanazi, O. R. Shahin et al., "Machine algorithm for heartbeat monitoring and arrhythmia detection based on ECG systems," Computational Intelligence and Neuroscience, vol. 2021, 2021.
[11]. A., O. R. S. El-Komy, m Rasha, M. A. El Aziz, and A. I. Taloba, "Integration of computer vision and natural language processing in multimedia robotics application," Information Sciences Letters, vol. 7, p. 6, 2022.
[12]. T. Hong, J. A. Choi, K. Lim, and P. Kim, "Enhancing personalized ads using interest category classification of SNS users based on deep neural networks," Sensors, vol. 21, no. 1, p. 199, 2020.
[13]. H. Yang, Q. Liu, S. Zhou, and Y. Luo, "A spam filtering method based on multi-modal fusion," Applied Sciences, vol. 9, no. 6, p. 1152, 2019.
[14]. B. Santoso, "An analysis of spam email detection performance assessment using machine learning," Jurnal Online Informatika, vol. 4, no. 1, pp. 53–56, 2019.
[15]. A. J. Karim, Shanmugam, Azam, Kannoorpatti, Jonkman, and Boer, "An intelligent spam detection model based on artificial immune system," Information, vol. 10, no. 6, p. 209, Jun. 2019.
[16] Ma, Thae Ma, Kunihito Yamamori, and Aye Thida. "A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification." In 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), pp. 324-326. IEEE, 2020.
[17]. Bhagyashri U. Gaikwad and P. P. Halkarnikar, " Spam E-mail Detection by Random Forests Algorithm", ISSN (Print): 2278-5140, Volume-2, Issue – 4, 2013.
[18]. Niclas Englesson, "Logistic Regression for Spam Filtering", Mathematical Statistics Stockholm University Bachelor Thesis 2016:9 http://www.math.su.se.