Sakarya University Journal of Science

ISSN : 2147-835X
Publisher : Sakarya University

Vol. 28, No. 2, 270-282, 2024
DOI: https://doi.org/10.16984/saufenbilder.1264476

Research Article

Machine Learning Based Classification for Spam Detection

Serkan Keskin[1*], Onur Sevli[2]

[1] Burdur Mehmet Akif Ersoy University, Institute of Science and Technology, Department of Computer Engineering, Burdur, Türkiye, serkankeskin@isparta.edu.tr
[2] Burdur Mehmet Akif Ersoy University, Faculty of Engineering and Architecture, Department of Computer Engineering, Burdur, Türkiye, onursevli@mehmetakif.edu.tr
* Corresponding author

ARTICLE INFO

Keywords:
Artificial Intelligence
Email Classification
Machine Learning
Spam Detection

Article History:
Received: 13.03.2023
Accepted: 08.12.2023
Online Available: 22.04.2024

ABSTRACT

Electronic Electronic messages, i.e. e-mails, are a communication tool frequently used by individuals or organizations. While e-mail is extremely practical to use, it is necessary to consider its vulnerabilities. Spam e-mails are unsolicited messages created to promote a product or service, often sent frequently. It is very important to classify incoming e-mails in order to protect against malware that can be transmitted via e-mail and to reduce possible unwanted consequences. Spam email classification is the process of identifying and distinguishing spam emails from legitimate emails. This classification can be done through various methods such as keyword filtering, machine learning algorithms and image recognition. The goal of spam email classification is to prevent unwanted and potentially harmful emails from reaching the user's inbox. In this study, Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms are used to classify spam emails and the results are compared. Algorithms with different approaches were used to determine the best solution for the problem. 5558 spam and non-spam e-mails were analyzed and the performance of the algorithms was reported in terms of accuracy, precision, sensitivity and F1-Score metrics. The most successful result was obtained with the RF algorithm with an accuracy of 98.83%. In this study, high success was achieved by classifying spam emails with machine learning algorithms. In addition, it has been proved by experimental studies that better results are obtained than similar studies in the literature.

1. Introduction

With the widespread use of the Internet, electronic communication has become more preferred. One of the most important tools of electronic communication is electronic messages, which we call e-mail. Today, individuals or organizations have one or more e-mail accounts. Instant delivery of messages, no cost and ease of use increase the importance and prevalence of e-mail [1]. According to Statista Research Department data, the number of actively used e-mail accounts in 2020 is more than 4 billion. This number is estimated to increase to 4.6 billion in 2025. In 2020, 306 billion e-mails are sent and received every day, and this number is expected to exceed 376 billion in 2025 [2].

The use of e-mail is not only practical but also has various vulnerabilities. The e-mail account to be hijacked in various ways, for e-mails containing advertisements etc. to hijack your computer by installing a software on your computer when you click on the advertisement, and for the installed software to disrupt communication by sometimes filling the

Cite as: S. Keskin, O. Sevli (2024). Machine Learning Based Classification for Spam Detection, *Sakarya University Journal of Science,* 28(2), 270-282. https://doi.org/10.16984/saufenbilder.1264476

bandwidth. Such unsolicited e-mails are characterized as "spam". Between October 2020 and September 2021, the global daily spam volume peaked in July 2021 with approximately 283 billion spam emails out of a total of 336.41 billion emails. By August 2021, this number had fallen to 65.50 billion. By September, the average spam volume had again increased by 36 percent, reaching 88.88 billion out of a total of 105.67 billion emails sent worldwide [3].
Email providers are expected to stop spam emails before they reach users. Many email providers include mechanisms that attempt to filter spam by comparing the sender address of emails against so-called blacklists of known spammers. However, since spammers frequently change their sender addresses, the success of these programs has not reached the desired level [4]. At this point, a more effective and flexible solution is needed. Generally, spam e-mails contain messages such as "easy money", "adult entertainment", etc. in their headers or content, which can deceive individuals. The process of classifying emails by interpreting messages is based on the keyword detection rule. This method has made the inadequacy of address-based filtering of spam e-mails more successful with keyword detection algorithms. Machine learning techniques, which have recently gained popularity and are used in many different fields, provide alternative solutions for filtering spam e-mails much more successfully.

1.1. Methods used to detect spam emails

Unsolicited emails (spam) are usually fake emails sent for advertising or fraudulent purposes and often contain content that users do not want or are not interested in. Such emails can put users in difficult situations or reduce work efficiency. Therefore, it is important to detect and filter spam emails.

1.1.1. Traditional spam detection systems

Such spam detection systems, which are not based on artificial intelligence, usually use simple algorithms that distinguish spam based on the content of the message, the sender's address or the content of its links. The effectiveness and accuracy of these systems is lower than that of AI-based systems. They are less flexible and adaptive than AI-based systems. The main methods used in traditional spam detection systems are as follows:

- Email authentication: This method is used to verify who the sender of an email is. It verifies the authenticity of the sender using standards such as DomainKeys Identified Mail (DKIM) and Sender Policy Framework (SPF). This makes it possible to detect fake emails or spam emails sent from fake accounts [5].

- List of email addresses: This method enables the detection of spam emails using a predefined list of email addresses. This list may include email addresses with a high probability of spam [6]. This method can be effective in preventing spam emails, but it also involves the risk of false positives, i.e., correct email addresses being falsely flagged as spam.

- Content filtering: This method is used to detect spam emails based on the content in the emails. For example, words and phrases such as advertisements, product sales or illegal content can be detected in emails and these emails can be marked as spam. This method can be effective in preventing spam emails, but it also involves the risk of false positives [7].

- Sharing a list of email addresses: This method enables the detection of spam emails by sharing a list of spam email addresses between different users and organizations. In this way, it enables the detection of spam emails by sharing a list of spam email addresses between different users and organizations [7].

1.1.2. Artificial intelligence-based spam detection systems

Artificial intelligence-based spam detection systems are software used to detect spam messages that are common in electronic communication networks. These systems use various artificial intelligence techniques to search for and detect specific characteristics of spam messages. Spam messages are usually marketing messages with a high content of advertisements and promotions. These messages are often sent to many people and are often unsolicited or unnecessary. Sending too many

spam messages wastes the time and effort of email users. Artificial intelligence-based spam detection systems are designed to reduce these problems. These systems examine the content, headers and other features of e-mail messages and classify spam messages according to certain criteria [8].

- Systems based on biological intelligence: Systems based on biological intelligence are artificial intelligence systems that mimic the structure and functioning of the human brain. Such systems have a high degree of adaptive and learning capabilities, mimicking the learning, remembering and problem-solving abilities of the human brain. In particular, they have a network structure that transmits signals from inputs to outputs using structures called neural networks. These neural networks can have learning and adaptive properties, much like the human brain. By mimicking the natural structure and functioning of the human brain, such systems can have a very high degree of adaptive and learning capabilities [9].

- Machine learning-based systems: Machine learning-based spam systems are systems that help to automatically detect spam emails. These systems usually identify spam emails using features such as keywords and phrases found in the content of the emails. They also take into account that spam emails are usually sent regularly and that they fit a certain profile of email addresses and domains used. Spam systems developed using machine learning learn from pre-labeled datasets and discover which features in these datasets are more effective in identifying spam emails [10].

- These features may include keywords and phrases in the content of the emails, the sender's email address and domain, the email header, and the format of the email. The learned features are used to detect spam emails and new incoming emails are evaluated according to these features. The advantages of machine learning-based spam systems are that they have high detection rates as they learn from pre-labeled datasets [11]. Furthermore, these systems can improve themselves through dynamic learning processes and become more accurate classifiers over time. However, the disadvantages of machine learning-based spam systems include errors such as decreasing correct detection rates if the datasets are not large and diverse enough, or mistakenly identifying non-spam emails as spam [12].

2. Literature Review

When we examine the studies conducted in the literature using artificial intelligence techniques for the detection of spam e-mails, it is seen that e-porta classification processes are performed with different algorithms. Some of these studies used traditional machine learning algorithms, while others used algorithms inspired by biological systems such as Artificial Neural Networks (ANN).

In a study classifying comments in different languages obtained from social media, an accuracy of 96% was achieved using the Naive Bayes (NB) algorithm [13]. In another study to classify e-mails, a dataset containing 5574 English messages was classified with 95.48% accuracy using the NB algorithm and 97.83% accuracy using the Support Vector Machine (SVM) algorithm [14]. In another study for filtering short messages (SMS), unwanted advertisements were tried to be distinguished. The highest scores obtained in the classification process were reported as 98.61% with SVM and 97.55% with NB [15].

In some studies, classification is performed with messages sent via social media. In the result obtained by classifying 1383 tweets, the accuracy rate of RF was 92.95% [16]. The same algorithm may not always be more successful in the results found. This is because different data sets are used. For example, in another spam e-mail detection study, 600 e-mails were classified. As a result of this classification, Naive Bayes was 95.5% and SVM was 93.5% [17]. In another study, 6000 emails were classified and Naive Bayes was 94.6% and SVM was 98.5% successful [18]. Another of the algorithms examined is LR. In this study, LR was used to classify incoming emails as raw and spam. Dedekurt et al. presented a new spam approach by combining LR and artificial bee colony [19].

In another study, the ABC-LR algorithm was more successful than the classical LR algorithm [20]. Janez-Martino used the LR algorithm on a spam dataset to evaluate the combination of LR with a bag of words [21]. Apart from this, it has been observed that certain algorithms such as Naive Bayes-based and SVM have been used more than other machine learning algorithms [22].
It was revealed that the NB algorithm was 96.31% successful in the classification of 310 e-mails using the similar word suggestion feature of the Zemberek library [23]. In a study conducted on 4327 mail data sets with simulated neural networks (SNN), the success rate was found to be 95.82 [24]. In a study with the nearest neighbor (KNN) algorithm, the highest success rate of the KNN algorithm was 97.50% on a dataset of 4601 e-mails taken from the UCI machine learning repository website [25]. In another study on the same data set, the SVM algorithm was 93.07% successful.

In the study conducted by Jain et al. they used a data set consisting of 5572 messages labelled as raw and spam. As a result of the classification, they achieved a success rate of 98.79% with the SVM algorithm [26]. On the same data set, Gadde et al. used the LSTM model and achieved a success rate of 98.5%. TF-IDF and Hashing Vectoriser were used in the model [27]. Reddy and Reddy achieved 95.32% success rate by using SVM algorithm on 5572 spam sms dataset [28]. In another study, 98.56% success rate was achieved by using NB algorithm [29]. In the study conducted by Abayomi et al. on the same data set, a 98.6% success rate was obtained with the BILSTM model using deep learning method. [30].

3. Material and Method

In this study, a classification study was carried out on a data set consisting of 5558 samples for distinguishing spam e-mails. After natural language processing, the results of the classifications performed with 5 different machine learning algorithms consisting of Random Forest, Logistic Regression, Naive Bayes, Support Vector Machine and Artificial Neural Network are reported in terms of different metrics.

3.1. Data set

In this study, a dataset consisting of 5558 samples and two attributes was used to detect spam e-mails. The first attribute is the English content text of the email message and the second attribute is the target label that indicates whether the email is spam or not. This csv file (spam.csv, 480.13 kB) prepared by Faisal Qureshi, contains 5558 unique instances of ham (87%) and spam (13%) messages. [31].

Of the instances in the dataset, 747 are marked as spam and 4811 are marked as non-spam. The graph showing the class distribution rates in the dataset is given in Figure 1.

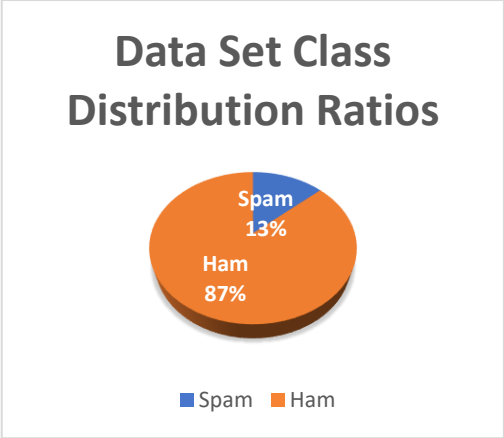

Figure 1. Class distribution rates

When the class distribution rates are analyzed, it is seen that the data set is not balanced. For this reason, cross-validation was applied in classification processes and detailed measurement metrics obtained through complexity matrices are reported.

3.2. Natural language processing (NLP)

Natural Language Processing (NLP) enables computers to communicate and process data using natural language. It is a sub-branch that uses technologies such as artificial intelligence and machine learning and typically works with text and audio data. NLP is artificial intelligence technologies that give humans the ability to understand and use natural language. NLP is divided into two main parts: text processing and audio processing.

Text processing works with text data and performs operations such as reading,

understanding and summarizing texts. Voice processing, on the other hand, works with voice data and performs operations such as recognizing voices, generating text from voices and translating texts into voice. In recent years, there has been a rapid development of NLP in phenomena such as question answering, machine translation and machine reading comprehension. NLP can be divided into three parts: modeling, learning and reasoning [32]. TF-IDF (Term Frequency-Inverse Document Frequency) is a natural language processing technique used to measure word importance in texts. TF-IDF calculates how often a word occurs in a text (Term Frequency, TF) and how few texts containing that word occur in total texts (Inverse Document Frequency, IDF). The product of these two values indicates the importance of the word. TF-IDF is used to better understand the meaning of texts. TF-IDF is widely used for measuring word distributions in texts and can be used in applications such as determining the similarity of texts, classifying texts or making connections between texts [33].

Each word in the dataset used in this study is associated with a numerical index value and those that carry spam flags are labeled. During the model training, the textual expressions in the dataset were separated word by word and subjected to numerical transformations, making it a completely numerical dataset. The dataset was classified with 5 different machine learning algorithms. In the study carried out with algorithms written in Python programming language in a spyder environment, tests were carried out using various library structures. With the algorithms applied to the dataset, performance evaluations were made according to precision, sensitivity, accuracy and F1 scores. All algorithms were subjected to 5-fold cross-validation.

3.3. Classification algorithms used

The data set used in the study was classified using 5 different machine learning algorithms: Support Vector Machine, Logistic Regression, Naive Bayes, Random Forest and Artificial Neural Network.

3.3.1. Support vector machine (SVM)

SVM is widely used in many studies because it produces significant accuracy with less computational power. SVM is one of the most popular supervised learning algorithms used to solve regression and classification problems. The goal of the SVM algorithm is to construct the best line or decision boundary that can classify data points in a multidimensional space that classifies them distinctly [34]. This boundary is called the hyperplane. The SVM selects endpoints or vectors to form the hyperplane. This selected state is called the support vectors [35]. The SVM algorithm is used in many different fields such as image classification, text classification and face detection.

3.3.2. Logistic regression (LR)

LR, like SVM, is one of the important machine learning algorithms among the algorithms that use supervised learning techniques. It is used to predict a categorical dependent variable using a set of independently given variables. LR predicts the output of a categorical dependent variable. It should give a discrete or categorical value as a result. The result can be true or false, 0 or 1. Instead of giving an exact value, it gives a probabilistic value between 0 and 1. Instead of a linear line, LR draws an "S" shaped function to cover two maximum values. This function curve gives the probability of whether a state exists or not [36]. LR is a highly successful machine learning algorithm that calculates probabilities using discrete and continuous data and classifies newly entered data.

3.3.3. Naive bayes (NB)

It is the first filtering algorithm used as a probabilistic classifier [37]. The NB algorithm is a supervised learning algorithm for solving classification problems based on Bayes theory. It is used for text classification with a high-dimensional training data set. The NB algorithm can make predictions quickly. It makes predictions by calculating the probability of the object. Due to their simplicity and high performance, these approaches are the most widely used in open-source systems proposed for spam filtering [38]. This algorithm is also used in

areas such as article classification and sentiment analysis.

3.3.4. Random forest (RF)

The RF algorithm is a machine learning algorithm created by combining many decision trees. This algorithm can be used for classification and regression problems. The RF algorithm is a combination of many decision tree models, each trained with different subsets of data. Each decision tree makes decisions on specific features and data points using a set of decision tree nodes. Decision trees work by dividing the data into small subsets and classifying the data points in these subsets with a set of decision nodes. [39].

This algorithm allows each decision tree to make predictions individually and eventually produces a result by combining all the predictions. This improves accuracy and consistency, giving better results than a single decision tree. A large number of trees in the forest provides higher accuracy [40]. Training time is less compared to other algorithms. It can maintain accuracy even if a certain part of the data is missing. It is generally used in banking, medicine, land use and marketing sectors.

3.3.5. Artificial neural network (ANN)

An Artificial Neural Network (ANN) is a machine learning model that works like the brain. Like a network of nerve cells in the brain, an ANN is made up of many nerve cells (neurons). Neurons are connected and process information by sending signals to each other. The ANN learns by using the connections between neurons and adjusting their weights [41]. Information is transmitted to the network from the input layer. It is then processed in the intermediate layer and sent to the output layer. The information coming into the network is converted into output using the weight value of the network. To produce the correct outputs, the evaluation of the weights must be done correctly. The process in ANN is to calculate the parameters w (weight) and b (bias) that will give the model the best score. [42]. ANN is a method that offers successful solutions to many problems we encounter in daily life such as classification, prediction and modeling.

3.4. Model performance measurement

A confusion matrix was used to express the performance of the classifier used. The confusion matrix is a table used to evaluate how well a class is distinguished from each other. It allows us to see how well the algorithm can predict the correct class. The rows of the matrix represent the predicted class and the columns represent the true class [43]. For a binary classification problem where the classes are "positive" and "negative", the general structure of the complexity matrix looks like Figure 2.

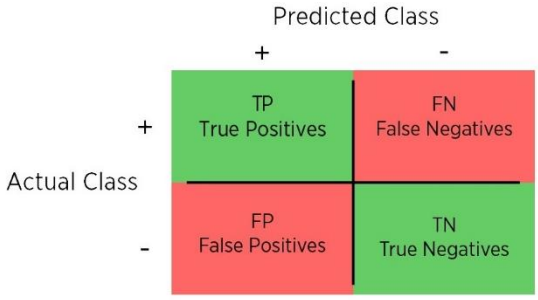

Figure 2. Complexity matrix

In machine learning, true positive refers to the number of correct positive predictions made by a model out of all positive predictions. In other words, it is the number of instances where the model correctly identifies a positive instance as positive. True negative refers to the number of correct negative predictions made by a model out of all negative predictions. It is the number of instances where the model correctly identifies a negative instance as negative. False positive refers to the number of false positive predictions made by a model out of all negative predictions. In other words, it is the number of instances where the model predicts a positive instance when it is negative. In machine learning, false negative refers to the number of false negative predictions made by a model out of all positive predictions. It is the number of instances where the model predicts a negative pattern when it is positive.

Different evaluation metrics can be calculated from a complexity matrix. These metrics are useful for understanding the performance of a classification algorithm and comparing the performance of different models. The formulas for deriving these measures from the complexity matrix are given in Table 1.

Table 1. Formulation of measurements

Measure	Description	Formula
Accuracy	Overall performance of model	$\frac{TP + TN}{TP + TN + FP + I}$
Precision	How accurate the positive predictions are	$\frac{TP}{TP + FP}$
Sensitivity	Coverage of actual positive sample	$\frac{TP}{TP + FN}$
F1 Score	Hybrid metric useful for unbalanced classes	$\frac{2TP}{2TP + FP + FN}$

Accuracy: The proportion of correct predictions. It is calculated as the number of true positives divided by the total number of true negatives divided by the number of predictions.

Precision: The proportion of correct positive predictions. It is calculated by dividing the number of true positives by the total number of true positives and false positives.

Sensitivity: The proportion of true positive cases that are correctly predicted. It is calculated by dividing the number of true positives by the total number of true positives and false negatives.

F1 Score: The harmonic mean of the precision and recall values. The F1 score takes values between 0 and 1, with higher values indicating better classification performance.

4. Experimental Study and Findings

In the study conducted for spam detection, the dataset consisting of 5558 samples was classified using 5 different machine learning algorithms: SVM, LR, NB, RF and ANN. Before the classification process, the e-mail message texts in the dataset were subjected to natural language processing. The texts were first parsed into sentences and then segmented into words according to the determined brackets. Word vectors were created and Term Frequency / Inverse Document Frequency was calculated. The mathematically transformed e-mail messages were classified with the specified algorithms using 5-fold cross-validation. The averages of the measurements obtained with each algorithm are reported.

The complexity matrix obtained as a result of the classification process performed with the SVM algorithm is given in Figure 3.

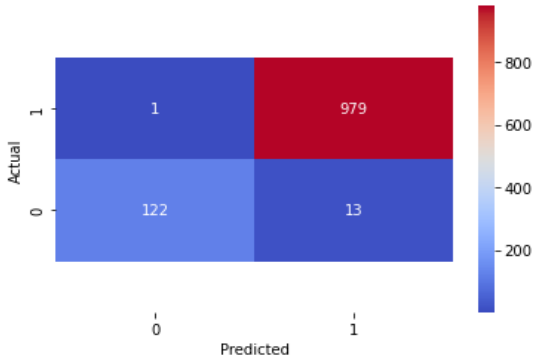

Figure 3. SVM results

In the complexity matrix of the DVM algorithm, it is understood that the model distinguishes between spam and non-spam emails with overall success. The values of the metrics calculated over the complexity matrix of the model are given in Table 2.

Table 2. Calculated metrics for SVM

SVM Metrics	Ratios
Accuracy	98.74
Precision	98.86
Sensitivity	99.89
F1 Score	99.29

In Table 2, the accuracy value showing the overall success of the model is 98.74%. The precision and sensitivity values showing the discrimination of the classes were obtained as 98.86% and 99.89%. The F1 Score value, which expresses the balance of these two values, was obtained as 99.29%.

The complexity matrix obtained as a result of the classification process performed with the LR algorithm is given in Figure 4.

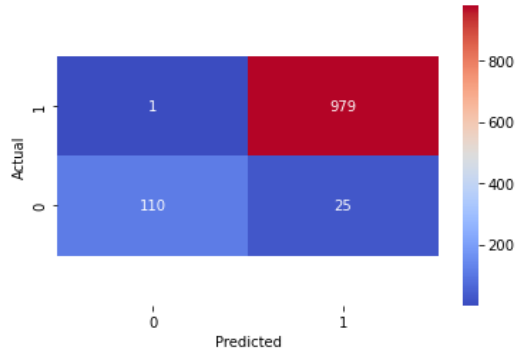

Figure 4. LR results

In the complexity matrix of the LR algorithm, it turns out that the model distinguishes spam and non-spam emails with general success. The values of the metrics calculated over the complexity matrix of the model are given in Table 3.

Table 3. Calculated metrics for LR

LR Metrics	Ratios
Accuracy	97.66
Precision	97.75
Sensitivity	99.89
F1 Score	98.68

In Table 3, the accuracy value showing the overall success of the model is 97.66%. The precision and sensitivity values showing the discrimination of the classes were obtained as 97.75% and 99.89%. The F1 Score value, which expresses the balance of these two values, was obtained as 98.68%.

The complexity matrix obtained as a result of the classification process performed with the NB algorithm is given in Figure 5.

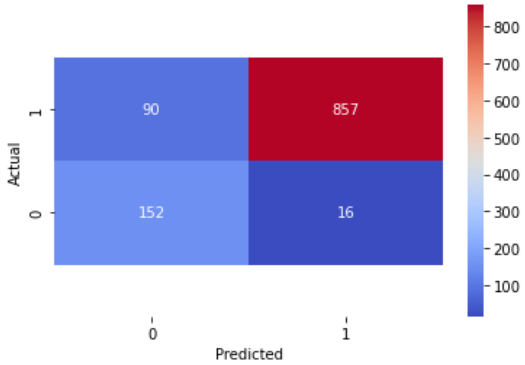

Figure 5. NB results

In the complexity matrix of the NB algorithm, it is understood that the model mixes TN and TP values with FN. This affects the success of the model. The values of the metrics calculated over the complexity matrix of the model are given in Table 4.

Table 4. Calculated metrics for NB

NB Metrics	Ratios
Accuracy	90.49
Precision	98.16
Sensitivity	90.49
F1 Score	94.17

In Table 4, the accuracy value showing the overall success of the model is 90.49%. The precision and sensitivity values showing the discrimination of the classes were obtained as 98.16% and 90.49%. The F1 Score value, which expresses the balance of these two values, was obtained as 94.17%.

The complexity matrix obtained as a result of the classification process performed with the RF algorithm is given in Figure 6.

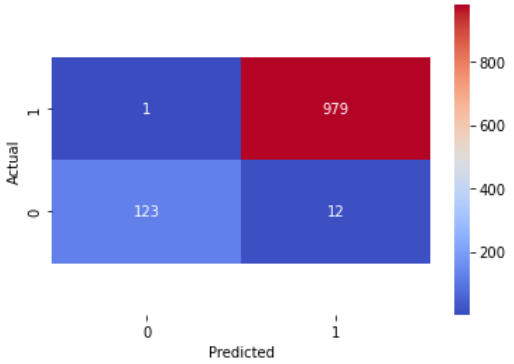

Figure 6. RF results

In the complexity matrix of the RF algorithm, it appears that the model distinguishes spam and non-spam emails with overall success. The values of the metrics calculated over the complexity matrix of the model are given in Table 5.

Table 5. Calculated metrics for RF

RF Metrics	Ratios
Accuracy	98.83
Precision	98.78
Sensitivity	99.89
F1 Score	99.34

In Table 5, the accuracy value showing the overall success of the model is 98.83%. The precision and sensitivity values showing the discrimination of the classes were obtained as 98.78% and 99.89%. The F1 Score value, which

expresses the balance of these two values, was obtained as 99.34%.

The complexity matrix obtained as a result of the classification process performed with the ANN algorithm is given in Figure 7.

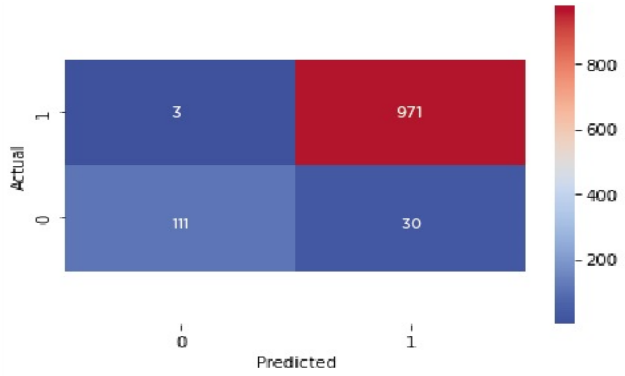

Figure 7. ANN results

In the Complexity matrix of the ANN algorithm, it is understood that the model successfully distinguishes between spam and non-spam emails in general. The values of the metrics calculated over the complexity matrix of the model are given in Table 6.

Table 6. Calculated metrics for ANN

ANN Metrics	Ratios
Accuracy	97.04
Precision	97.00
Sensitivity	99.69
F1 Score	98.32

In Table 6, the accuracy value showing the overall success of the model is 97.04%. The precision and sensitivity values showing the discrimination of the classes were obtained as 97.00% and 99.69%. The F1 Score value, which expresses the balance of these two values, was obtained as 98.32%.

The measurements obtained as a result of the classification processes performed with 5 different algorithms are summarized in Table 7.

Table 7. Calculated measurements of the algorithms used

Machine Learning Algorithm	Accuracy	Precision	Sensitivity	F1 Score
SVM	98.74	98.86	99.89	99.29
LR	97.66	97.75	99.89	98.68
NB	90.49	98.16	90.49	94.17
RF	98.83	98.78	99.89	99.34
ANN	97.04	97.00	99.69	98.32

When Table 7, which shows the classification performance of the algorithms, is analysed, it is revealed that the RO algorithm ranks first with 98.83% accuracy in terms of overall success. The NB algorithm showed the lowest performance with 90.49% accuracy. In terms of F1 score, which expresses the balance in distinguishing the classes, the most successful algorithm was RO with 99.34%, while the lowest success was NB algorithm with 94.17%. It is understood that RO>DVM>LR> ANN> in the general success ranking.

The comparison of the findings obtained in the classification process performed in this study with other similar studies in the literature is given in Table 8. In this table, the most successful algorithm and accuracy rates are given.

The last row in Table 8 is the result of this study. The reason why the accuracy rates in some studies in this table are close to the accuracy rates of our study is that the data set sizes and data sets are close to each other. As it can be understood, it has been experimentally demonstrated that this study is more successful than other studies. This is due to the fact that the natural language processing processes of the study are more successful than other similar studies.

5. Conclusion

E-mail is one of the most widely used communication tools and one of the biggest problems in the use of this tool is spam messages. Spam messages are e-mails that are intended to advertise or deceive and their detection is of great importance. Various techniques and algorithms have been proposed to detect spam e-mails.

In the present study, 5 different machine learning algorithms were used to classify spam e-mails using a dataset of 5558 samples consisting of spam and non-spam e-mail messages. With 5-fold cross-validation, the results of the classification processes are reported with accuracy, precision, sensitivity and f1 score metrics.

In the study, the rf algorithm produced the most successful result with 98.83% accuracy. In this study, unlike other studies, the use of natural language processing made the success different and high. It is concluded that this score is higher than similar studies in the literature. This study sets an example for a machine learning-based infrastructure that will consistently filter spam content in e-mail servers. In future studies, it is aimed to obtain higher performance results with different algorithms on datasets to be prepared for different natural languages**.**

Table 8. Comparison table of the most successful accuracy rates on the same and different data sets

Study Name	Data Set Used	Most Successful Algorithm	Highest Accuracy (%)
Kumar and al., 2023 [29]	Spam Dataset	NB	98.56
Jain and al., 2022 [26]	Spam Dataset	SVM	98.79
Abayomi and al., 2022 [30]	Spam Dataset	BILSTM	98.60
Reddy and Reddy, 2021 [28]	Spam Dataset	SVM	95.32
Gadde and al., 2021 [27]	Spam Dataset	LSTM	98.50
Junnarkar and al., 2021 [4]	Data set containing 5574 e-mails	SVM	97.83
Ma and al., 2020 [21]	6000 data sets containing e-mails	SVM	95.5
Salihi, 2019 [16]	1183 units obtained from Twitter the resulting data set	RF	92.95
Karamollaoglu and Dogru, 2018 [6]	TurkishMail dataset consisting of 600 e-mails	NB	95.5
Nazlı, 2018 [44].	Data set consisting of 300 e-mails	SVM	98.33
Kale, 2018 [45]	Data set of 4,709 e-mails	Gradient Boosted Tree (GBT)	94.97
Yıldız, 2017 [31]	Data set of 310 Turkish e-mails	NB	96.31
Alkaht and al., 2016 [28]	CSDMC 2010, SpamAssassin, Tarassul	SNN	95.82
Sharma and Suryawanshi, 2016 [29]	Spambase	KNN	97.50
Zavvar al., 2016. [46]	Spambase	SVM	93.07
This study	**Spam Dataset**	**SVM 98.74** **LR 97.66** **NB 90.49** **RF 98.83** **ANN 97.04**	**98.83**

Article Information Form

Funding

The author (s) has no received any financial support for the research, authorship or publication of this study.

Authors' Contribution

All authors have contributed in experimental study and writing of the manuscript equally.

The Declaration of Conflict of Interest/ Common Interest

No conflict of interest or common interest has been declared by the authors.

The Declaration of Ethics Committee Approval

This study does not require ethics committee permission or any special permission.

The Declaration of Research and Publication Ethics

The authors of the paper declare that they comply with the scientific, ethical and quotation rules of saujs in all processes of the paper and that they do not make any falsification on the data collected. In addition, they declare that sakarya university journal of science and its editorial board have no responsibility for any ethical violations that may be encountered, and that this study has not been evaluated in any academic publication environment other than sakarya university journal of science.

References

[1] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi, & O. E. Ajibuwa, “Machine learning for email spam filtering: review, approaches and open research problems.”Heliyon, 5(6), e01802, 2019.

[2] L.Ceci (2022, Nov. 14). Number of e-mail users worldwide [online]. Available:https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/

[3] S. Dixon (2022, Apr. 28) Daily spam volume worldwide Available: https://www.statista.com/statistics/1270424/daily-spam-volume-global/

[4] P.Pantel, D. L. Spamcop, "A Spam Classification and Organization Program." Learning for Text Categorization, 2006.

[5] S. Zeadally, E. Adi, Z. Baig, & I. A. Khan, "Harnessing artificial intelligence capabilities to improve cybersecurity." Ieee Access 8, 23817-23837, 2020.

[6] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, & M. Alazab, "A comprehensive survey for intelligent spam email detection." IEEE Access 7, 168261-168295, 2019.

[7] T. Dogan, "On Term Weighting for Spam SMS Filtering." Sakarya University Journal of Computer and Information Sciences 3.3, 239-249, 2020.

[8] S. Douzi, F. A. AlShahwan, M. Lemoudden, & B. El Ouahidi, "Hybrid email spam detection model using artificial intelligence." International Journal of Machine Learning and Computing 10.2 2020.

[9] E. M. Onyema, S. Dalal, C. A. T. Romero, B. Seth, P. Young, & M. A. Wajid, "Design of intrusion detection system based on cyborg intelligence for security of cloud network traffic of smart cities." Journal of Cloud Computing 11.1, 1-20, 2022.

[10] A. Bhowmick, S. M. Hazarika, "E-mail spam filtering: a review of techniques and trends." Advances in Electronics, Communication and Computing: ETAEERE-2016, 583-590, 2018.

[11] D. Abidin, The Effect of Derived Features on Art Genre Classification with Machine Learning. Sakarya University Journal of Science, 25(6), 1275-1286, 2021

[12] P. Sharma, U. Bhardwaj. "Machine learning based spam e-mail detection. "International Journal of Intelligent Engineering and Systems 11.3, 1-10, 2018

[13] Ö. Şahinaslan, H. Dalyan, E. Şahinaslan, "Naive bayes sınıflandırıcısı kullanılarak youtube verileri üzerinden çok dilli duygu analizi. "Bilişim Teknolojileri Dergisi 15.2, 221-229, 2022

[14] A. Junnarkar, S. Adhikari, J. Fagania, P. Chimurkar, D. Karia "E-mail spam classification via machine learning and natural language processing." 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). IEEE, 2021.

[15] Y. S. Bozan, Ö. Çoban, G. T. Özyer, & B. Özyer, "SMS spam filtering based on text classification and expert system." 2015 23nd Signal Processing and Communications Applications Conference (SIU). IEEE, 2015.

[16] A. K. A. Salihi, Spam detection by using word-vector learning algorithm in online social networks. MS thesis. Fen Bilimleri Enstitüsü, 2019.

[17] H. Karamollaoglu, İ. A. Dogru, M. Dorterler, "Detection of Spam E-mails with Machine Learning Methods. "2018 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, 2018.

[18] M. T. Ma, K. Yamamori, A. Thida, "A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification."2020 IEEE 9th Global Conference on Consumer Electronics (GCCE). IEEE, 2020.

[19] B. K. Dedeturk, B. Akay. "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. "Applied Soft Computing 91 106229, 2020.

[20] N. Baktır, A. Yılmaz, "Makine Öğrenmesi Yaklaşımlarının Spam-Mail Sınıflandırma Probleminde Karşılaştırmalı Analizi. "Bilişim Teknolojileri Dergisi 15.3: 349-364, 2022.

[21] F. Jánez-Martino, E. Fidalgo, S. González-Martínez, J. Velasco-Mata, "Classification of spam emails through hierarchical clustering and supervised learning. "arXiv preprint arXiv: 2005.08773, 2020.

[22] R. Mansoor, N. D. Jayasinghe, M. M. A. Muslam. "A comprehensive review on email spam classification using machine learning algorithms. "2021 International Conference on Information Networking (ICOIN). IEEE, 2021.

[23] A. Yıldız, M. Demirci, Kurumsal e-posta sınıflandırma sistemi. Diss. Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, 82, Ankara, 2017.

[24] I. J. Alkaht, B. Al-Khatib. "Filtering spam using several stages neural networks." Int. Rev. Comp. Softw 11.2, 2016.

[25] A. Sharma, A. Suryawanshi. "A novel method for detecting spam email using KNN classification with spearman correlation as distance measure. "International Journal of Computer Applications 136.6, 28-35, 2016

[26] Jain, T., Garg, P., Chalil, N., Sinha, A., Verma, V. K., & Gupta, R. SMS spam classification using machine learning techniques. In 2022 12th international conference on cloud computing, data science & engineering (confluence) (pp. 273-279). IEEE, 2022.

[27] Gadde, S., Lakshmanarao, A., & Satyanarayana, S. SMS spam detection using machine learning and deep learning techniques. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 358-362). IEEE, 2021.

[28] Reddy, G. A., & Reddy, B. I. Classification of Spam Text using SVM. Journal of University of Shanghai for Science and Technology, 23(8), 616-624, 2021

[29] Kumar, R., Murthy, K. S. R., Ramesh Babu, J., & Shaik, A. Live Text Analyzer to Detect Unsolicited Messages Using Count Vectorizer. Journal of Engineering Sciences, 14(06), 2023.

[30] Abayomi-Alli, O., Misra, S., & Abayomi-Alli, A. A deep learning method for

automatic SMS spam classification: Performance of learning algorithms on indigenous dataset. Concurrency and Computation: Practice and Experience, 34 (17), e6989, 2022.

[31] 'Email Spam Detection 98% Accuracy | Kaggle'. https://www.kaggle.com/code/mfaisalqureshi/email-spam-detection-98-accuracy/data (accessed Aug. 21, 2023).

[32] M. Zhou, N. Duan, S. Liu, H. Y. Shum, "Progress in neural NLP: modeling, learning, and reasoning."Engineering 6.3, 275-290, 2020.

[33] I. Yahav, O. Shehory, D. Schwartz, "Comments mining with TF-IDF: the inherent bias and its removal. "IEEE Transactions on Knowledge and Data Engineering 31.3, 437-450, 2018

[34] Y. Altuntaş, A. F. Kocamaz, A. M. Ülkgün, "Determination of Individual Investors' Financial Risk Tolerance by Machine Learning Methods. "2020 28th Signal Processing and Communications Applications Conference (SIU). IEEE, 2020.

[35] R. Gürfidan, M. Ersoy, "Classification of death related to heart failure by machine learning algorithms. "Advances in Artificial Intelligence Research 1.1, 13-18, 2021

[36] S. Şenel, B. Alatli. "Lojistik regresyon analizinin kullanıldığı makaleler üzerine bir inceleme. "Journal of Measurement and Evaluation in Education and Psychology 5.1, 35-52, 2014.

[37] A. McCallum, K. Nigam. "A comparison of event models for naive bayes text classification. "AAAI-98 workshop on learning for text categorization. Vol. 752. No. 1. 1998.

[38] V. Metsis, I. Androutsopoulos, G. Paliouras. "Spam filtering with naive bayes-which naive bayes?", CEAS. Vol. 17. 2006.

[39] F. M. Avcu, "Az Veri Setli Çalışmalarında Derin Öğrenme Ve Diğer Sınıflandırma Algoritmalarının Karşılaştırılması: Agonist Ve Antagonist Ligand Örneği "İnönü Üniversitesi Sağlık Hizmetleri Meslek Yüksek Okulu Dergisi 10.1, 356-371, 2022

[40] Ö. Akar, O. Güngör, "Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. "Jeodezi ve Jeoinformasyon Dergisi 106, 139-146, 2012.

[41] A. Arı, M. E. Berberler, "Yapay sinir ağları ile tahmin ve sınıflandırma problemlerinin çözümü için arayüz tasarımı. "Acta Infologica 1.2, 55-73, 2017

[42] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, A. M. Umar, O. U. Linus, M. U. Kiru, "Comprehensive review of artificial neural network applications to pattern recognition. "IEEE Access 7, 158820-158846, 2019

[43] Z. K. Şentürk, "Artificial neural networks based decision support system for the detection of diabetic retinopathy. "Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi 24.2, 424-431, 2020.

[44] N. Nazlı, Analysis of machine learning-based spam filtering techniques. MS thesis. 2018.

[45] B. Kale, Veri madenciliği sınıflandırma algoritmaları ile e-posta önemliliğinin belirlenmesi. MS thesis. Fen Bilimleri Enstitüsü, 2018.

[46] M. Zavvar, M. Rezaei, S. Garavand. "Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine. "International Journal of Modern Education and Computer Science 8.7, 68, 2016.