Avrupa Bilim ve Teknoloji Dergisi
Sayı 21, S. 690-696, Ocak 2021
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 21, pp. 690-696, January 2021
Copyright © 2021 EJOSAT
Research Article

Separation of Incoming E-Mails Through Artificial Intelligence Techniques

Mete Yağanoğlu[1*], Erdal Irmak[2]

[1*] Atatürk University, Faculty of Engineering, Department of Computer Engineering, Erzurum, Turkey, (ORCID: 0000-0003-3045-169X), yaganoglu@atauni.edu.tr
[2] Gazi University, Faculty of Technology, Department of Electrical and Electronics Engineering, Ankara, Turkey, (ORCID: 0000-0002-4712-6861), erdal@gazi.edu.tr

(First received 15 November 2020 and in final form 31 January 2021)

(**DOI:** 10.31590/ejosat.841299)

ATIF/REFERENCE: Yağanoğlu, M., & Irmak, E. (2021). Separation of Incoming E-Mails Through Artificial Intelligence Techniques. *European Journal of Science and Technology*, (21), 690-696.

Abstract

Technological developments are making individuals and organizations ever more dependent on e-mail to communicate and share information. The increasing use of e-mail as an essential and popular communication method poses potentially severe threats to the Internet and society. Spam e-mails cause security problems for internet users and waste storage, bandwidth, and productivity resources. The increase in the volume of spam e-mails has created an intense need to develop more reliable and robust antispam filters. Therefore, it has become necessary to recommend adaptive spam detection models. In this paper, an intelligent system for the detection and filtering of spam e-mails is described. Machine learning methods aim to create the best models using the available data and analyze new data most accurately, with the help of the model created using previous data. In this study, spam detection was carried out using machine learning methods. In this study, K-nearest neighbors, support vector machine, and decision trees were used in the classification stage. The classification achieved an accuracy of 98.2% in spam detection.

Keywords: Spam Detection, Natural Language Processing, Artificial Intelligence, Machine Learning.

Yapay Zeka Teknikleri İle Gelen E-Postaların Ayrıştırılması

Öz

Teknolojik gelişmeler, bireyleri ve kuruluşları, iletişim kurmak ve bilgi paylaşmak için e-postalara daha bağımlı hale getirmektedir. E-postaların internet üzerinden önemli ve popüler bir iletişim olarak artan kullanımı, İnternet'i ve toplumu etkileyen ciddi bir tehdit oluşturmaktadır. Spam e-postalar internet kullanıcıları için güvenlik sorunlarına sebep olmaktadır ve depolama, bant genişliği ve üretkenlik açısından kaynakları boşa harcamaktadır. İstenmeyen e-postaların hacmindeki artış, daha güvenilir ve sağlam antispam filtrelerin geliştirilmesi için yoğun bir ihtiyaç yaratmıştır. Bu nedenle, uyarlanabilir spam algılama modellerinin önerilmesi bir gereklilik haline gelmektedir. Bu çalışmada, spam e-postalarını başarılı bir şekilde tespit etmek ve filtrelemek için yapay zekaya dayalı akıllı bir algılama sistemi önerilmektedir. Makine öğrenimi yöntemleri, mevcut verileri kullanarak en iyi modelleri oluşturmayı ve önceki veriler kullanılarak oluşturulan model yardımıyla yeni verileri en doğru şekilde analiz etmeyi amaçlamaktadır. Bu çalışmada sınıflandırma aşamasında k-en yakın komşu, destek vektör makinesi ve karar ağaçları kullanılmıştır. Bu çalışmada, istenmeyen e-posta tespiti makine öğrenimi yöntemleri kullanılarak gerçekleştirilmiştir ve % 98.2 başarı oranına ulaşılmıştır.

Anahtar Kelimeler: Spam Tespiti, Doğal Dil İşleme, Yapay Zeka, Makine Öğrenmesi.

[*] Corresponding Author: yaganoglu@atauni.edu.tr

1. Introduction

An e-mail has become practical and popular for correspondence, as the number of Internet users has increased. The increasing use of the Internet and web technology has changed the way in which people use computers. Although the Internet is used for research, examinations, and entertainment, it also provides an environment in which users can share their ideas and get feedback. Millions of people use e-mails for personal, business, marketing, education, and other communication purposes. E-mail management is a significant and growing problem because it tends to be misused by individuals and organizations. Sending a large number of unsolicited bulk e-mails is called spam e-mail.

Spam detection is the subject of considerable research (Spirin & Han, 2012; Shi & Xie,2013; Sirivianos et al., 2011; Khamis et al., 2020; Al-Ajeli et al., 2020). Spam has become a platform of choice used by cybercriminals to spread malicious loads such as viruses and trojans (AlMahmoud et al., 2017). For these purposes, e-mails must be separated and forwarded to users.

Spam filtering is a typical two-class problem involving separating legitimate messages from spam. Spam filtering separates incoming e-mails into spam or ham. It prevents spam from coming to a user’s e-mail without the user seeing it. Separation of spam before it arrives in the mailbox is the most critical step for spam filtering. Spam is often described as unsolicited or unsolicited bulk electronic messages. The rapid increase in the amount of spam makes it increasingly difficult to filter e-mail manually.

There are many spam filtering techniques (Dada et al., 2019). Contextual Filtering Methods consist of automatic filtering rules and detect incoming e-mails using classification approaches. Contextual filtering was developed to separate spam e-mails by evaluating the words and phrases in the e-mail using different analytical methods. Spam and ham e-mails are removed from the user’s mailbox in the Case-Based Spam Filtering method. A machine learning algorithm is used to train datasets and to test whether incoming mail is spam. The Intuitive Spam Filtering Method uses pre-built rules to evaluate many patterns, which are usually regular expressions, against a selected message (Dada et al., 2019; Christina et al.,2010). The Previous Similarity-Based Spam Filtering Method approach uses sample-based machine learning methods to classify incoming e-mails according to their similarity to stored samples. This approach uses the k-closes running algorithm to filter spam e-mails (Sakkis et al., 2001). Adaptive spam filtering classifies incoming spam e-mails by separating them using client-based filter management (Dada et al., 2019; Pelletier et al., 2004)

2. Related Works

In the study by Asghar et al., a dataset from Amazon’s website and sentences tagged for spam detection was used, and spam detection was done (Asghar et al.,2020).

Tan et al. propose a community decision approach that combines the characteristics of e-mails to detect spam effectively (Tan et al., 2018).

Tekerek proposes a spam SMS detection technique was proposed using Data Mining methods. A dataset containing 747 spam SMS and 4827 ham SMS was used. Cross-validation technique was used to evaluate the spam SMS estimation in the dataset. The proposed study achieved a 98,33% success rate for the Support Vector Machine (SVM) algorithm (Tederek, 2019).

Gunawan et al. performed a correct classification process with a success rate of 96.49% in their spam detection study using 985 text messages, 860 spam, and 125 non-spam (Gunawan et al., 2018).

El-Alfy et al. proposed a model for filtering messages for e-mail. They analyzed various methods to conclude the features that were determined so that the complexity could be reduced. The authors used features such as SVM and Naive Bayes algorithms and URLs, spam domain, defect words, recipient address, and subject area (El-Alfy et al., 2016).

Faris et al. proposed a detection system using an automatic identification feature to separate spam e-mails. Experimental results were obtained as the proposed system's accuracy was 92.2%, recall 97.6%, and precision 93.3% (Faris et al., 2019).

AlMahmoud et al. proposed Spamdoop, a common spam detection platform facility that protects big data privacy. Spamdoop uses a fairly parallel coding technique that allows spam campaigns to be detected at competitive times (AlMahmoud et al., 2017).

Saleh et al. provide a study on the detection of anomalies in spam e-mails. In their studies, it was observed that performance continued to improve with the inclusion of more datasets, and a true 98.5% spam and raw detection rate increased, while the Real Positive and Real Negative detection rate increased by 6% (Saleh et al., 2019).

Zhu and Tan proposed a method to extract attributes for spam e-mails. In the method, a two-dimensional feature is created by estimating the spam and non-spam e-mail concentrations. Then, all the properties of each field are combined into one feature vector. Various experiments were carried out on four comparison companies using 10-fold cross-validation. This approach has been shown to be able to extract information about effective location from messages (Zhu, Y., & Tan,2010).

Olatunji proposed an SVM-based model for spam detection. For training and test sets, 95.87% and 94.06% accuracy were obtained, respectively (Olatunji, 2019).

In the study by Kumar et al., Hidden Markov Model and ID3 were used to identify e-mails as spam or raw. For this purpose, an Enron dataset of 5172 e-mails containing 2086 Spam and 2086 raw pre-classified e-mails was used, achieving a success rate of 89% (Kumar et al.,2018).

Given the work is done so far and the accuracy of performance achieved to date, it is clear that more research is needed about the possibility of getting better results using the same popular datasets. Therefore, this study is set up to create an alternative model that can push the accuracy level to a higher level than previous models.

3. Material and Methods

Machine learning methods aim to create the best model using the available data and to analyze the new data most accurately with the help of the model created with the previous data as new data arrive. Machine learning and artificial intelligence methods have been frequently used recently to separate spam e-mails successfully. In this study, spam detection was made by machine learning methods.

In the studies conducted, classification methods have been proposed to separate the incoming e-mails. These techniques calculate the rate of occurrence of keywords or patterns in e-mail messages and decide properties (Dada et al., 2019).

3.1. Natural Language Processing (NLP)

NLP, a sub-science of artificial intelligence and linguistics, is defined as sensing texts in natural languages and sound waves by computer, analyzing software, and transferring to the computer environment. Natural language processing is the study of training human language computationally. In other words, it is the science of teaching computers how to understand and produce human language. NLP aims to increase the comprehension and understanding skills of the computer by processing the language that people use by speaking and writing in daily life. It aims to briefly establish a semantic link between human language and computers (Yao, 2019). In this study, NLP has been used for sentence segmentation.

3.2. Natural Language Tool Kit (NLTK)

NLTK was first established as part of the computer science course in 2001. NLTK includes adaptive and calculable linguistic modules written as open source. It stands for natural language toolkit. It is an open-source library created with over 50 corpus and lexical resources developed and developed with Python programming language to work with human language data. There are also several modules in this library; these modules are the packages we will use while pre-processing our data, using machine learning algorithms, processing with the Twitter API, etc. For instance, tokenization in a sentence and stemming operations by removing the existing attachments in the Word (Yao, 2019). In this study, NLTK was used in pre-processing, tokenization, and stemming stages. NLTK also saves us from dealing with unnecessary words during the pre-processing phase of a data set, that is when we are going to make the data the machine can understand.

3.3. Spam Filtering Process

Spam filtering aims to minimize the volume of spam e-mails. Filtering is the process of separating harmful e-mails to detect malicious applications and eliminate the effects of this. Spam filters are distributed in front of the e-mail server or in the mail relay with the firewall (Katakis et al., 2007; Liu & Gouda, 2008). E-mail senders can forward e-mails to a mail server that processes e-mails for many clients on the Internet. The mail server can use a spam filter to remove spam and then forward the filtered e-mails to addressed clients. Filters can be applied to clients, where they can be installed, and computers to mediate between some endpoint devices (Christina et al.,2010).

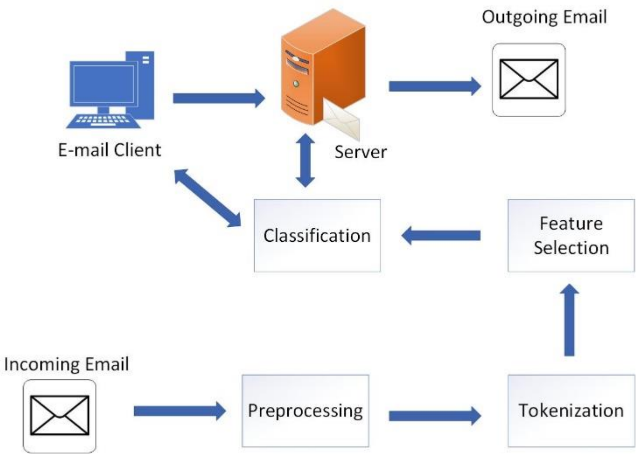


Figure 1. Spam filtering architecture

Figure 1 shows the steps for spam filtering. Incoming mail is divided into training and test models after pre-processing and feature extraction. The label of the e-mail received during the training is evident. After the training phase is completed, the e-mail received during the test phase is separated.

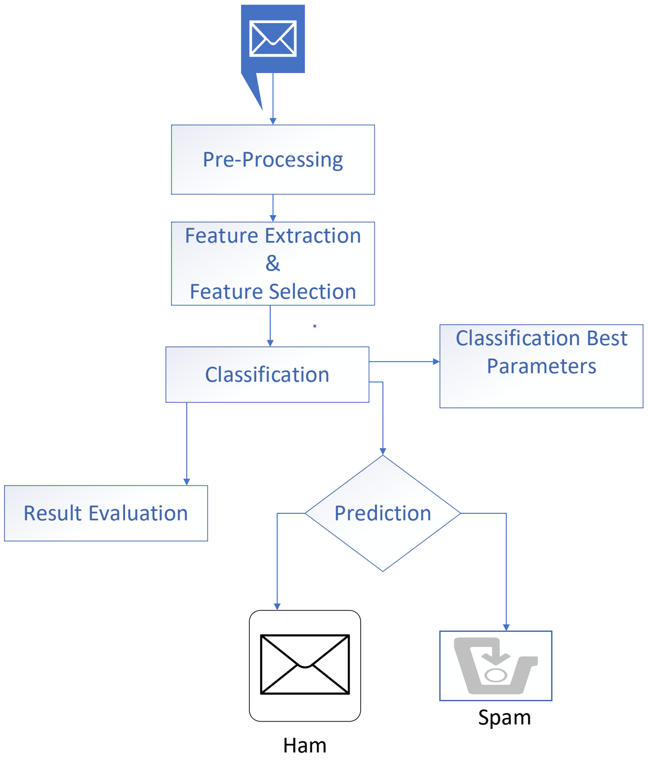

Pre-Processing
Feature Extraction & Feature Selection
Classification
Classification Best Parameters
Result Evaluation
Prediction
Ham
Spam

Figure 2. The steps for spam filtering

The necessary steps to be observed in the mining of data in an e-mail message can be divided into the following categories: These are pre-processing, feature extraction, feature selection, and classification. Figure 2 shows the basic steps in the proposed spam filtering framework. It creates a classification model during the training phase, using a group of pre-classified messages. During this phase, messages are pre-processed and analyzed to extract related features. A vector represents each message. Verification can be performed on-demand while the classifier is being trained. Once a classifier is created, it will be deployed to estimate the class of newly received messages.

3.3.1 Pre-processing

The use of e-mails in daily life causes the number of spam e-mails to increase. For this reason, spam e-mails need to be separated successfully. One of the most important steps of successful spam filtering is the implementation of appropriate pre-processing steps.

As a result of these operations, the link texts in the e-mails and the links of the e-mails were distinguished and pre-processed. Thanks to the pre-processing stage, all of the link texts were capitalized, the problem of foreign characters was resolved, punctuation was removed, some words that contain some special meaning, and words shorter than three letters were distinguished.

Pre-processing is the method used and developed to obtain meaningful data from the incoming e-mail. These require some standard Natural Language Processing (NLP) pre-processing steps such as uppercase, lowercase conversion, noise removal, lexicon normalization, object standardization, clipping word suffixes, stemming, lemmatization, and frequency of terms. It is necessary to prepare the e-mails ready for analysis. These pre-processing steps can affect the overall performance of the detection algorithm. Before spam filtering, the content of the e-mail is marked by dividing each word of the e-mail and then saving them in a word list data structure. For this purpose, the python library was used to learn.

3.3.2 Feature Extraction & Feature Selection

After the pre-processing, the content of the e-mails is saved in a list as a list of words containing all relevant words that each classified e-mail has. Then, every e-mail in the dataset is checked again, and if the word contained in the e-mail is in the word list, then it is labeled according to the classified e-mail, which is raw or spam. For instance, if the word “click” is present in the word list and is found with a Spam e-mail and the result is positive, the same word in the list is labeled spam. In other words, it is separated by adding it to the list so that it is spam or non-spam.

3.3.3 Classification

Classified data is the process of classification of unclassified data using various algorithms. K-nearest neighbors (KNN) is a nonparametric and simple learning algorithm. KNN is used in data mining, attack detection systems in providing information security, in many areas of genetics and bioinformatics, and in many similar systems such as pattern recognition systems. The smallest K is determined depending on the number of K among the ordered values. The neighboring sample closest to the sample to be tested is determined. Class labels of K neighbors found for the classification of the sample to be tested are used. Choosing the most appropriate value for K is done by examining the data. KNN protects all training data and makes decisions based on the training data set. KNN is based on determining the distances between an unknown object and each training set object (Deng et al., 2016).

SVM is an algorithm used because it gives a high success rate to solve classification and regression problems. One of the simple and highly effective classification methods used in classification problems is support vector machines. SVM, one of the statistical learning algorithms and developed by Vapnik, has yielded successful results in many real problems. SVM is a controlled classification algorithm based on statistical learning theory. SVM is known as the training algorithm based on the probability distribution of statistical techniques. The SVM working principle aims to maximize the vertical distance of these samples to the separating plane, in other words, the hyper-plane by finding the closest samples of the classes while classifying the data. The basic logic of SVM is to determine the best separating plane for data structures that can be separated linearly. Thus, the misclassification of the data in both the training and test set was minimized. In a linear separable situation, there can be many decision planes that separate classes. SVM detects the greatest distance between the two classes from these planes. The vectors closest to this plane are also called support vectors. In nonlinear problems, samples are moved to a space where they can be separated in higher dimensions and linearly, and the solution is made in this new space (Torabi et al., 2015).

Decision tree (DT) is a machine learning algorithm that produces successful results for spam filtering. DT, class-known sample data is divided into small groups with simple decision-making steps. Data similar to each other are grouped with each division process, and classification is made by induction method. DT should make relatively little effort from users during the training of datasets. It is very useful because the decision trees can be easily applied to very large and missing datasets and both continuous and categorical variables can be analyzed because the results are understandable. DT is a nonparametric method that is an alternative to the least-squares and logistic regression method and does not include the necessary assumptions for regression-type problems (Dada et al., 2019).

After our dataset is separated as training and test, pre-processing, feature extraction, and classification will be done and the incoming e-mail will be separated. E-mails must be classified and tagged according to the categories of ham e-mail or spam e-mail. All features are used for spam and ham detection during testing. If these words are later found to be spam, they are compared to the duplicate set of properties when saved as spam, and the e-mail data is converted into two values, spam and ham. Our classifier then classifies this e-mail as spam if spam is more likely than ham. In this way, test e-mails are classified as ham or spam.

4. Experimental Results

4.1. Dataset

Spam e-mails have been used randomly as datasets in the literature (Almaida et al., 2011; Hidalgo et al. 2012). In Dataframe, the pandas library of the dataset, there are two columns of data, with 5572 object-type data in each column (Almaida et al., 2011). The ratio of ham and spam data is shown in Figure 3 (Number of spam: 4825, number of ham: 747). After

30% of the dataset was reserved for the test, pre-processing, feature selection, and classification steps were applied.

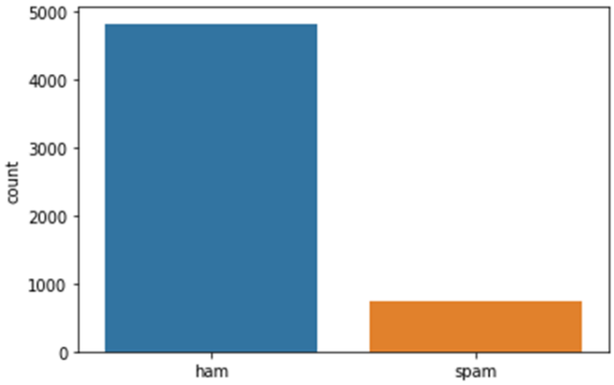


Figure 3. Dataset spam/ham ratio

4.2. Evaluation Criterion

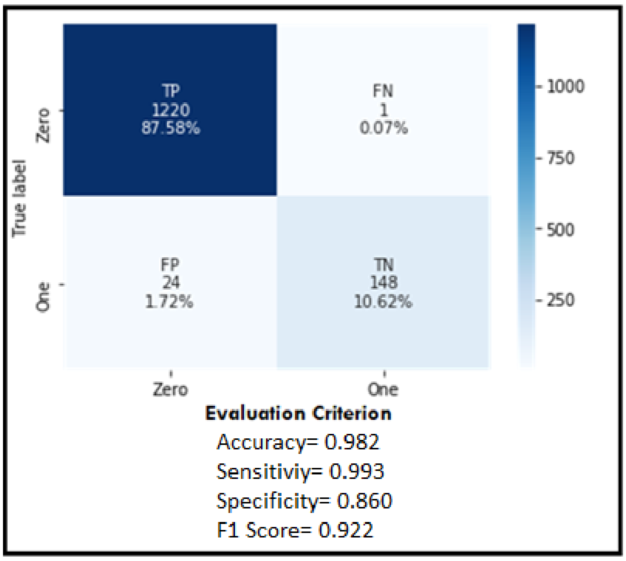
Confusion matrix was used to measure model performance. True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN) figures are given in the confusion matrix (Bozkurt et al., 2020; Yağanoğlu and Köse, 2018). According to these figures, accuracy, sensitivity, specificity, and F1 score were calculated.

4.3. Success Rates

After our dataset was separated into training and test datasets, pre-processing steps such as tokenizing, removing unnecessary words, rooting, finding sentence elements, removing structures in the sentence, and feature selection were applied. In the classification step, three different classifiers were used. The results of the decision tree classification, confusion matrix, and evaluation are shown in Figure 4.

As seen in Figure 4, spam e-mails were identified with an accuracy of 98.2%. The figure shows TP, FN, FP, and TN values and their ratios. Also, the accuracy, sensitivity, specificity, and F1 score values were calculated and are shown. The results of the KNN classification, confusion matrix, and evaluation are shown in Figure 5.

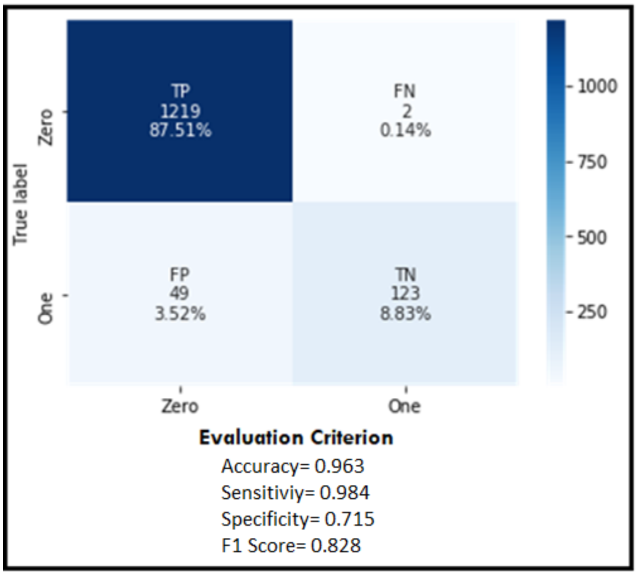
As seen in Figure 5, spam e-mails were identified with an accuracy of 96.3%. The figure shows TP, FN, FP, and TN values and their ratios. Also, the accuracy, sensitivity, specificity, and F1 score values were calculated and are shown. The results of the SVM classification, confusion matrix, and evaluation are shown in Figure 6.



Evaluation Criterion

Accuracy= 0.982
Sensitiviy= 0.993
Specificity= 0.860
F1 Score= 0.922

Figure 4. Decision Tree Confusion Matrix and Evaluation Criterion



Evaluation Criterion

Accuracy= 0.963
Sensitiviy= 0.984
Specificity= 0.715
F1 Score= 0.828

Figure 5. KNN Confusion Matrix and Evaluation Criterion

As seen in Figure 6, spam e-mails were identified with an accuracy of 96.7%. The figure shows TP, FN, FP, and TN values and their ratios. Also, the accuracy, sensitivity, specificity, and F1 score values were calculated and are shown.

Classification results are as shown in Table 1. As seen from the table, the best success rate was determined by the Decision Tree algorithm.

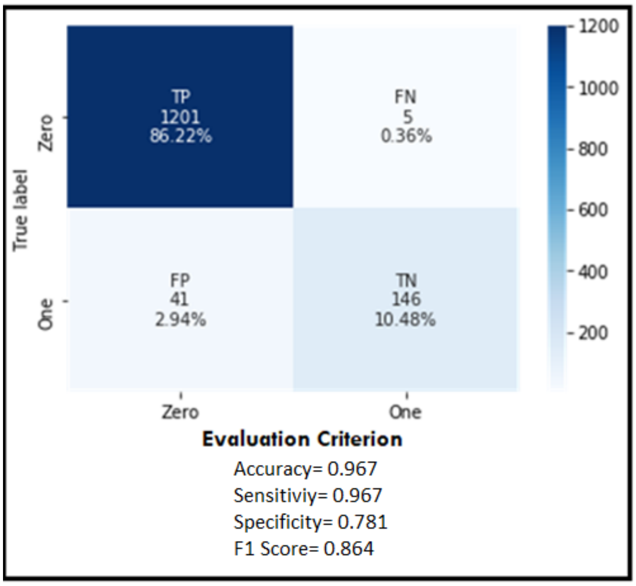


Figure 6. SVM Confusion Matrix and Evaluation Criterion

Tablo 1. Classification Results

Classifier	Accuracy	Sensitivity	Specificity	F1 Score
DT	0.982	0.993	0.860	0.922
SVM	0.963	0.984	0.715	0.828
KNN	0.967	0.967	0.781	0.864

5. Discussion and Conclusions

In this study, artificial intelligence techniques were applied to the spam filtering problem. After application of the pre-processing and feature extraction methods, the data were classified as ham or spam to classify incoming e-mails. Attempts by different researchers to solve the spam problem using machine learning classifiers are discussed. The architecture of e-mail spam filters and the processes for filtering spam e-mails are examined. Public datasets and performance metrics that can be used to measure the effectiveness of spam filters were investigated. The spam threat of machine learning algorithms was effectively addressed, and comparative studies of existing classification techniques were carried out.

Datasets in the literature containing messages labeled as spam or non-spam were used to determine the study's success. Some of the data were used as training data and the rest as a query dataset. The data were pre-processed and trained using a Vector Space Model. An accuracy of 98.2% in spam detection was achieved.

The dramatic increase in spam in recent years has created considerable interest among many researchers. There has been significant progress in the field of spam filtering. Spam e-mail is a common type of cyber-problem that all Internet users encounter in their daily lives. Spam e-mails waste resources and pose serious security threats. Detection and filtering are still the most appropriate solutions to combatting spam e-mails. Almost all e-mail servers run some types of spam e-mail filters on incoming e-mails, but we all have firsthand experience with the frustration of spam e-mails, and we are constantly experiencing it. The biggest problem is to immediately identify new types of spam e-mails with no information beforehand. Spammers constantly and quickly adopt new techniques to bypass spam filters and continue to create new types of spam e-mail. Most spam e-mail filters require some information about the nature of spam e-mails, and detection is often difficult. For these reasons, more research is needed to increase the effectiveness of spam filters. This research will facilitate the development of spam filters using machine learning approaches. We hope that researchers will use this study to conduct qualitative research in spam filtering using transfer learning and deep learning algorithms.

References

Al-Ajeli, A., Alubady, R., & Al-Shamery, E. S. “Improving spam e-mail detection using hybrid feature selection and sequential minimal optimization”. Indonesian Journal of Electrical Engineering and Computer Science, 19(1), 535-542, 2020.

AlMahmoud, A., Damiani, E., Otrok, H., & Al-Hammadi, Y. “Spamdoop: A privacy-preserving Big Data platform for collaborative spam detection”. IEEE Transactions on Big Data, 2017.

Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. “Contributions to the study of SMS spam filtering: new collection and results”. In Proceedings of the 11th ACM symposium on Document engineering, pp. 259-262, 2011.

Asghar, M. Z., Ullah, A., Ahmad, S., & Khan, A. “Opinion spam detection framework using hybrid classification scheme”. Soft computing, 24(5), 3475-3498, 2020.

Bozkurt, F., Köse, C., & Sarı, A. “A texture-based 3D region growing approach for segmentation of ICA through the skull base in CTA”. Multimedia Tools and Applications, *79*(43), 33253-33278, 2020.

Christina, V., Karpagavalli, S., & Suganya, G. “E-mail spam filtering using supervised machine learning techniques”. International Journal on Computer Science and Engineering (IJCSE), 2(09), 3126-3129, 2010.

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

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. “Efficient kNN classification algorithm for big data”. Neurocomputing, 195, 143-148, 2016.

El-Alfy, E. S. M., & AlHasan, A. A. “Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm”. Future Generation Computer Systems, 64, 98-107, 2016.

Faris, H., Ala'M, A. Z., Heidari, A. A., Aljarah, I., Mafarja, M., Hassonah, M. A., & Fujita, H. "An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks". Information Fusion, 48, 67-83, 2019.

Gunawan, D., Rahmat, R. F., Putra, A., & Pasha, M. F. "Filtering Spam Text Messages by Using Twitter-LDA Algorithm". IEEE International Conference on Communication, Networks and Satellite (Comnetsat), pp. 1-6, IEEE, 2018.

Hidalgo, J. M. G., Almeida, T. A., & Yamakami, A. "On the validity of a new SMS spam collection". 11th International Conference on Machine Learning and Applications, Vol. 2, pp. 240-245, IEEE, 2012.

Katakis, I., Tsoumakas, G., & Vlahavas, I., E-mail mining: Emerging techniques for e-mail management. In Web Data Management Practices: Emerging Techniques and Technologies (pp. 220-243). IGI Global, 2007.

Khamis, S. A., Foozy, C. F. M., Ab Aziz, M. F., & Rahim, N. "Header Based E-mail Spam Detection Framework Using Support Vector Machine (SVM) Technique". In International Conference on Soft Computing and Data Mining, pp. 57-65,. Springer, Cham,2020.

Kumar, V., Kumar, P., & Sharma, A. "Spam E-mail Detection using ID3 Algorithm and Hidden Markov Model". In 2018 Conference on Information and Communication Technology (CICT) (pp. 1-6). IEEE, 2018.

Liu, A. X., & Gouda, M. G. "Diverse firewall design. IEEE Transactions on Parallel and Distributed Systems". 19(9), 1237-1251, 2008.

Olatunji, S. O. "Improved e-mail spam detection model based on support vector machines". Neural Computing and Applications, 31(3), 691-699, 2019.

Pelletier, L., Almhana, J., & Choulakian, V. "Adaptive filtering of spam". In Proceedings. Second Annual Conference on Communication Networks and Services Research, pp. 218-224, IEEE, 2004.

Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C. D., & Stamatopoulos, P. "Stacking classifiers for antispam filtering of e-mail". arXiv preprint cs/0106040., 2001.

Saleh, A. J., Karim, A., Shanmugam, B., Azam, S., Kannoorpatti, K., Jonkman, M., & Boer, F. D. "An intelligent spam detection model based on artificial immune system". Information, 10(6), 209, 2019.

Shi, W., & Xie, M. "A reputation-based collaborative approach for spam filtering". AASRI Procedia, 5, 220-227,2013.

Sirivianos, M., Kim, K., & Yang, X. "Socialfilter: Introducing social trust to collaborative spam mitigation". In 2011 Proceedings IEEE INFOCOM, pp. 2300-2308, IEEE,2011.

Spirin, N., & Han, J. "Survey on web spam detection: principles and algorithms". ACM SIGKDD explorations newsletter, 13(2), 50-64,2012.

Tan, Y., Wang, Q., & Mi, G. "Ensemble decision for spam detection using term space partition approach". IEEE transactions on cybernetics, 50(1), 297-309, 2018.

Tekerek, A. "Support vector machine based spam SMS detection". Politeknik Dergisi, 22(3), 779-784,2019.

Torabi, Z. S., Nadimi-Shahraki, M. H., & Nabiollahi, A. "Efficient support vector machines for spam detection: a survey". International Journal of Computer Science and Information Security, 13(1), 11,2015.

Yağanoğlu, M., & Köse, C. "Real-time detection of important sounds with a wearable vibration based device for hearing-impaired people". Electronics, *7*(4), 50, 2018.

Yao, J. "Automated Sentiment Analysis of Text Data with NLTK". In Journal of Physics: Conference Series (Vol. 1187, No. 5, p. 052020). IOP Publishing, 2019.

Zhu, Y., & Tan, Y. "Extracting discriminative information from e-mail for spam detection inspired by immune system". In IEEE Congress on Evolutionary Computation (pp. 1-7). IEEE, 2010.