Batman Üniversitesi Yaşam Bilimleri Dergisi
Batman University Journal of Life Sciences
E-ISSN: 2459-0614
DergiPark AKADEMİK

Batman Üniversitesi Yaşam Bilimleri Dergisi 14 (1), 2024, 43-56

A Comparative Analysis of Learning Techniques in the Context of Turkish Spam*

Öznur ŞENGEL[1]*

*[1]Istanbul Kültür University, TURKEY

Doi: 10.55024/buyasambid.1501609

ARTICLE INFO

Article Type: Research Article

Article history:
Received: 15.06.2024
Received in revised form:
Accepted: 25.06.2024
Available online: 07.07.2024

Keywords:
Turkish SMS Datasets, Spam SMS Detection, SMS Classification, Machine Learning, Deep Learning.

*[1]Öznur ŞENGEL
E-mail address:
ozsengelnur@gmail.com
Orcid: 0000-0002-2186-927X

ABSTRACT

Short Message Service (SMS) is a mobile messaging tool used by billions of people to communicate via a mobile phone. However, due to the lack of proper message filtering techniques, this form of communication is vulnerable to unwanted and junk messages. This paper compared SMS spam detection approaches based on machine learning methods such as Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Multinominal Naïve Bayes (MNB), Logistic Regression (LR), and Support Vector Machines (SVM) and deep learning methods such as Convolutional Neural Networks (CNNs), Artificial Neural Networks (ANNs), and Long Short Term Memory (LSTM) in terms of f-score, accuracy, recall, precision, and a confusion matrix constructed for each strategy. The study tested two different preprocessing methods on two different Turkish SMS datasets to evaluate the approaches. The aim of this study is to contribute to the issue of spam filtering in Turkey. The results indicate that the highest accuracy values were achieved with Support Vector Machine (99.03%) using the first preprocessing method and Logistic Regression and Random Forest (98.07%) using the second preprocessing method on the BigTurkishSMS dataset, a combination of the two datasets used. As is the case with the majority of machine learning algorithms, the second preprocessing of the data set yielded superior results in deep learning models. The ANN model achieved the highest accuracy, with a score of 97.41%. The study employed a comparison of machine learning and deep learning techniques on Turkish SMS datasets, which will provide valuable insights for researchers working in this field.

* "This article is derived from the paper titled 'A Comparative Analysis of Learning Techniques in the Context of Turkish Spam' presented at the International Information Congress 2024 (IIC2024) held at Batman University between May 2-4, 2024."

Türkçe Spam Tespiti Bağlamında Öğrenme Tekniklerinin Karşılaştırmalı Analizi

Öznur ŞENGEL[1]*

*[1]İstanbul Kültür Üniversitesi, TÜRKİYE

Doi: 10.55024/buyasambid.1501609

MAKALE BİLGİSİ

Makale Türü: Araştırma Makalesi

Makale Geçmişi:
İlk gönderim tarihi: 15.06.2024
Düzeltme tarihi:
Kabul tarihi: 25.06.2024
Yayın tarihi: 07.07.2024

Anahatar Kelimeler:
Türkçe SMS Veri Kümeler, İstenmeyen SMS Tespiti, SMS Sınıflandırma,Makine Öğrenmesi, Derin Öğrenme.

*[1]Öznur ŞENGEL
E-mail address: ozsengelnur@gmail.com
Orcid: 0000-0002-2186-927X

ÖZET

Kısa Mesaj Servisi (SMS), milyarlarca insan tarafından cep telefonu aracılığıyla iletişim kurmak için kullanılan bir mobil mesajlaşma aracıdır. Ancak, uygun mesaj filtreleme tekniklerinin eksikliği nedeniyle, bu iletişim biçimi istenmeyen ve önemsiz mesajlara karşı savunmasızdır. Bu makalede, Adaptif Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), K-En Yakın Komşular (KNN), Karar Ağacı (DT), Rastgele Orman (RF), Multinominal Naïve Bayes (MNB), Lojistik Regresyon (LR) ve Destek Vektör Makineleri (DVM) gibi makine öğrenimi yöntemleri ile Evrişimli Sinir Ağları (CNN), Yapay Sinir Ağları (YSA) ve Uzun Kısa Süreli Bellek (LSTM) gibi derin öğrenme yöntemlerine dayalı SMS spam tespit yaklaşımları f-skor, doğruluk, duyarlılık, kesinlik ve her bir strateji için oluşturulan karışıklık matrisi açısından karşılaştırılmıştır. Çalışma, yöntemleri değerlendirmek için iki farklı ön işleme yöntemini iki farklı Türkçe SMS veri kümesi üzerinde test etmiştir. Bu çalışmanın amacı, Türkiye'deki spam filtreleme konusuna katkıda bulunmaktır. Sonuçlar, kullanılan iki veri kümesinin bir kombinasyonu olan BigTurkishSMS veri kümesi üzerinde en yüksek doğruluk değerlerinin birinci ön işleme yöntemi kullanılarak Destek Vektör Makinesi (%99,03) ve ikinci ön işleme yöntemi kullanılarak Lojistik Regresyon ve Rastgele Orman (%98,07) ile elde edildiğini göstermektedir. Makine öğrenimi algoritmalarının çoğunda olduğu gibi, veri setinin ikinci ön işlemesi derin öğrenme modellerinde üstün sonuçlar vermiştir. YSA modeli %97,41'lik bir skorla en yüksek doğruluğu elde etmiştir. Bu çalışma, Türkçe SMS veri kümeleri üzerinde makine öğrenimi ve derin öğrenme tekniklerinin bir karşılaştırmasını yaparak bu alanda çalışan araştırmacılar için değerli bilgiler sağlamaktadır.

1. INTRODUCTION

The United Nations World Population Prospects report (2022) states that the current global population is 8.08 billion, having increased by 74 million people since last year, representing a year-onyear growth of 0.9%. According to the latest data from GSMA Intelligence (Matthew, 2023), 69.4% of the world's population now uses a mobile device, with the global total increasing by 138 million (+2.5%) since the beginning of 2023. At the beginning of 2024, the number of mobile phone users was 5.61 billion (Kemp, 2024). By 2022, over 75% of Turkey's population will own a smartphone, with 68.7 million mobile internet users. At the Cyber Security Weekend - META event, Kaspersky reported a 120% increase in mobile threats in Turkey. The number of smartphone users in Turkey is expected to increase to approximately 88 million by 2028 (Dierks, 2023). It is predicted that the number of mobile phone users will lead to an increase in potential threats via mobile phones.

Short message service (SMS) is a mobile messaging tool that enables the exchange of 160-character text messages among mobile devices. Supported by almost all mobile devices, SMS provides

a quick and easy way to send messages to individuals or organizations for business and personal communication. Many companies that use Short Message Service, one of the main sources of communication, send SMS messages programmatically using Short Message Service Application Programming Interface (SMS API) software systems and send information messages to their customers. The SMS API, which has many advantages such as automation, integration, scalability, personalization, and real-time communication, is being abused by some users. During the day, mobile phone users are unintentionally exposed to unnecessary messages sent in this way.

Short Message Service (SMS) has security weaknesses, including phishing spam messages from cybercriminals. Cybercriminals are continuously enhancing their social engineering techniques to execute more effective phishing attacks. The topics of phishing attacks include traffic fines, bank transfer confirmations, overdue payments, money transfers, online orders, e-tickets, and similar items. Additionally, there are fraudulent messages that appear to originate from government offices, online shops, airlines, and booking services. It is important to remain vigilant and cautious when receiving messages from unknown sources. Criminals may also exploit holidays and major events. While there have been several studies on filtering and classifying spam messages in English (Salman et al., 2024), limited research has been conducted on spam detection in Turkish like Bengali (Al Maruf et al., 2023), and Indonesia (Theodorus et al., 2021). This study aims to fill this gap by classifying Turkish SMS messages as either spam or normal messages.

In this study, only two datasets containing spam messages in Turkish are used. The two datasets were pre-processed in two different ways. Finally, the two Turkish SMS datasets were combined to create a new, comprehensive dataset. In the new dataset, machine learning and deep learning techniques were evaluated for two different preprocessing methods. To compare the results, a confusion matrix was constructed for each technique and evaluated f-score, recall, precision, and accuracy metrics. The main contributions of this study are as follows:

- The performance of machine learning and deep learning techniques used in the literature is compared and evaluated using commonly used metrics.
- The techniques are tested on real Turkish SMS datasets.
- A new Turkish SMS dataset is obtained from two real datasets.
- Two different preprocessing methods are compared on real Turkish SMS datasets.
- Two real Turkish SMS datasets are used to filter spam messages.

In the remainder of this paper, Section 1.1 presents a review of past work in this area. Section 2 provides information about the datasets used, the data preprocessing steps applied, and details of the machine and deep learning techniques discussed in this paper. The test results of the techniques used in the datasets are presented in Section 3. Finally, the paper is discussed and concluded in Section 4.

1.1. Spam SMS Detection

Mobile devices are susceptible to a range of threats, which have increased over time and pose a significant risk to users. SMS attacks are a serious concern. The conventional BOTNET attack causes significant financial losses by surreptitiously sending SMS messages at regular intervals to toll lines that charge higher rates than regular phone lines (Masum and Samet, 2018). The transmission of unsolicited messages, commonly referred to as spam, can result in several adverse consequences. These include the dissemination of unwanted advertisements (Chen et al., 2022), the theft of personal data, the exposure to fraudulent schemes and commercial practices, and the installation of adware and hacking software.

Spam on SMS, WhatsApp, and other messaging services is a common danger that compromises phone security by spreading adware on phone devices. A phone device that sends spam may also be the result of a security breach. Mobile SMS spam is a nuisance to mobile device users and can harm the infrastructure of modern portable electronic devices, like email spam (Özdemir et al., 2013; Ergin and Isik, 2014a; Ergin and Isik, 2014b; Karamollaoglu et al., 2018; Eryılmaz et al., 2020).

The fact that the incoming message comes from a known and trusted source, that it contains information that the user expects and trusts from a previously agreed user, that he/she has received similar messages from the same person or organization before, and that he/she thinks it does not involve any risk, seems to be sufficient for the user to trust incoming messages and open them. If one or more of these criteria are met, the user opens the message and begins to analyze its content. In the meantime, he/she decides, based on experience, whether it is unnecessary or not and whether it is harmful or not. However, this experience can be costly. By generating unnecessary traffic on the mobile phone while the message arrives, it wastes the user's time and leaves them vulnerable to risks that may arise later. It is absolutely necessary to use a filter against such negative situations that may occur. The initial step in addressing the issue of unwanted messages is to implement a filtering system that intercepts messages as they enter and leave the phone. At this stage, the reliability of the previous sender's information, the quarantine list, the control of international spam lists, and the inclusion of some filters created with learning algorithms can be realized. Researchers have employed various learning techniques to filter out unwanted messages. With the richness of the dataset and powerful hardware, spam filtering has been developed using both deep learning models (Karasoy and Ballı, 2022) and machine learning methods (Srinivasarao and Sharaff, 2023).

Theodorus et al. (2021) developed a model that was trained with 4,125 messages and tested with 1,260 messages using a 10-fold cross-validation procedure to categorize spam, raw, and promotional messages in Indonesian dispatches. The results indicate that Multinomial Logistic Regression (MLR) (93.57%), XGBoost (93.52%), SVM (93.38%), and RF (93.62%) were effective classifiers. Arulprakash and Jansi (2021) divided the structure of SMS transmission into two layers and applied SMS spam detection filters and techniques to either the Access Layer or the Service Provider Layer. The accuracy values for grouping raw and spam messages using the NB, SVM, KNN, and DT algorithms were 95.2%, 88%, 85%, and 83%, respectively.

Message filtering is based on natural language processing, and the preprocessing steps used to create the training data set affect the success achieved. There are studies that create training datasets with various preprocessing steps and test them on application platforms such as Weka. Using the Orange 3 application, Özlem (2019) determined that the Neural Networks algorithm (98.4%) is highly accurate for the Turkish SMS dataset, while the NB algorithm (98.4%) is highly accurate for the UCI SMS Spam dataset. Suleiman et al. (2020) proposed a new message detection classifier supported by the H2O platform and employed three classifiers (RF, DL, and NB) and two validation models (3-fold and 10- fold cross-validation). The results indicate that the RF classifier has an accuracy rate of 0.977, making it the most effective. While NB has the lowest accuracy, it performed the best in terms of running time, with a value of 0.6 seconds.

Ertuğrul and Kaya (2016) proposed a new statistical approach that is superior to statistical methods that collect the frequency of occurrence of letters and words, called one-dimensional ternary models (1DTP), for analyzing text messages. The proposed method involves converting the text to UTF8 and comparing each letter with its adjacent letter. Two different feature sets were extracted from the results of these comparisons, and several machine learning techniques were tested on three different datasets. The accuracy rates obtained, 94.10%, 93.318%, and 87.15%, demonstrated the successful applicability of the proposed approach for message filtering feature extraction. The study conducted by Sajedi et al. (2016) aimed to test 15 different algorithms on the most common SMS datasets from 5 different countries available on the internet, and the Dendritic Cell Algorithm (DCA) achieved the highest accuracy rate of 99.95%.

Previous studies in the field of SMS spam detection have utilized a limited set of features, such as word frequency (Almeida et al., 2011; Mathew and Issac, 2011) and TF-IDF (Roy et al., 2020), as well as models like multiple model stack structures (Gupta et al., 2018) and various word embedding techniques (Jain et al., 2019). In this work, we use a Bag of Words (BoW) approach for word embedding, which is a feature extraction method that represents and describes the occurrence of words in a document.

2. METHODOLOGY

2.1. Datasets

The number of publicly available SMS datasets in the literature is limited, particularly for the Turkish language. Therefore, this study utilizes a pre-existing collection of SMS messages in Turkish, which is one of the most widely used suffixed languages worldwide. The first Turkish SMS dataset (Uysal et al., 2013) in the academic literature, named TurkishSMS, consists of 430 normal and 420 spam messages collected from different people. As seen in Figure 1, 50.6% of the messages are normal, and 49.4% are spam messages in the dataset. This dataset is the fundamental Turkish dataset used in articles written about SMS spam classification.

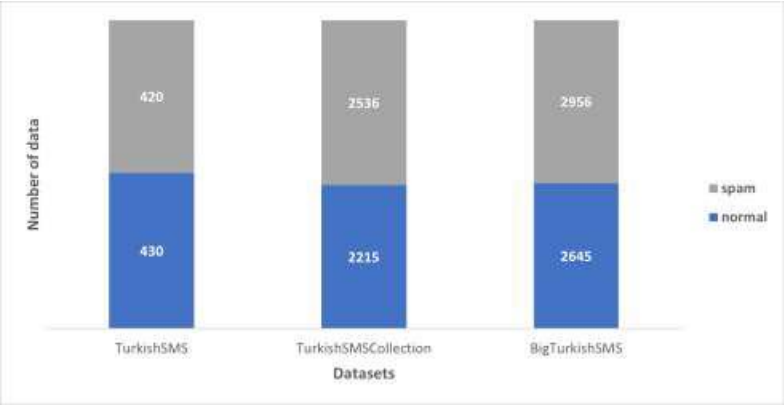


Figure 1. The number of spam and normal data in all used datasets

The TurkishSMSCollection (Karasoy and Ballı, 2022) dataset comprises messages collected from 76 individuals of varying age groups and residing in different cities across Turkey, including Muğla, İstanbul, and Ankara. The age groups represented were 18–24 (26 individuals), 25–39 (37 individuals), and over 40 (13 individuals). As seen in Figure 1, 46.6% of the messages are normal, and 53.4% are spam messages in the dataset.

As the TurkishSMS and TurkishSMSCollection datasets are both balanced, the experimental results obtained from these datasets can be compared fairly. The BigTurkishSMS dataset, which was created by combining the TurkishSMS and TurkishSMSCollection datasets, shows that 47.2% of the messages are normal and 52.8% are spam.

2.2. Data Preprocessing

Electronic messages are textual data and belong to the category of unstructured data. Although electronic messages can be read and viewed by computers, they need to be transformed into structured data sets by going through various processes in order to run machine learning algorithms on them. These procedures are necessary to convert the text from human language to a machine-readable format. The most important of these steps is preprocessing.

In our study, two preprocessing steps are applied. The steps applied in the first preprocessing step are as follows: (1) All letters in the message are converted to lowercase. (2) Abbreviations such as http, www, com, and tr in the text are deleted. (3) Emails in the text are deleted. (4) Numbers and spaces in the text are deleted. (5) Punctuation marks in the text are deleted.

The steps applied in the second preprocessing stage are as follows: (1) All special characters and punctuation marks in the message are deleted. (2) All letters in the message are converted to lowercase. (3) Stop words are deleted. (4) The bag of words was used to identify the 50 most frequently occurring words in the dataset.

2.3. Machine and Deep Learning

Machine learning (ML) is the ability to use the most appropriate algorithm to transform a dataset into model. Which ML method (supervised, unsupervised, etc.) works best is determined by the type of data being analyzed, the resources available, the nature of the data, and the intended outcome at the conclusion of the process. Deep learning is a sub-branch of machine learning that aims to learn phenomena through a nested hierarchy of concepts, mimicking the neurons of the human brain. In this study, Multinominal Naive Bayes (MNB), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Convolutional Neural Networks (CNNs), Artificial Neural Networks (ANNs), and Long Short-Term Memory (LSTM) algorithms are applied.

3. EXPERIMENTAL ANALYZES

3.1. Metrics

Nowadays, the accuracy results of many data science models used for classification processes are shared. However, it is not known to what extent the accuracy results reflect reality. Therefore, some techniques have been developed. In order to evaluate the performance of classification models used in machine learning, the confusion matrix, which compares the predictions and actual values of the target attribute, is frequently used.

Table 1. The confusion matrix

		Actual	
		True	**False**
Predicted	**True**	TP	FP
	False	FN	TN

The values represented on the confusion matrix (Table 1) are as follows:

• True Positive (TP) refers to cases that indicate a positive state and are predicted as such by the classifier.

• True Negative (TN) samples reflect a negative state and are predicted as such by the classifier.

• False Positives (FP) are samples that indicate a negative condition but are projected as positive by the classifier.

• False Negative (FN) refers to cases that exhibit a positive condition but are predicted as negative by the classifier.

Some values calculated using confusion matrix are used to evaluate classifier performance.

• Accuracy is the ratio of the number of correct predictions made by the classifier to the number of data in the whole data set. Accuracy, which is calculated as shown in Equation (1), measures how often the classifier makes a correct prediction.

$$TP + TN / TP + FN + FP + FN \quad (1)$$

• Precision is obtained by dividing the correct positive predictions made in all classes by the total positive predictions. Precision measures how accurately a prediction is made. Precision is also referred to as a positive predictive value and is calculated as shown in Equation (2).

$$Precision = TP / (TP + FP) \quad (2)$$

• Recall is determined as the number of accurate positive predictions divided by the total of the number of correct positive and negative predictions by using Equation (3).

$$Recall = TP / (TP + FN) \quad (3)$$

• The f-score is the harmonic mean of precision and recall, which is the ratio of true positive values. The f-score is a measure of how well a classifier performs. It is frequently used in literature to compare the performance of classifiers and is calculated as shown in Equation (4).

$$F\text{-score} = 2 \times ((Precision \times Recall) / Precision + Recall)) \quad (4)$$

3.2. Performance Analysis

Table 2 presents the recall, precision, accuracy, and f-score results obtained from training the TurkishSMS dataset with machine learning algorithms after the first and second preprocessing. The highest accuracy value of 95% was achieved by the AdaBoost and SVM algorithms, while KNN had the lowest performance with a value of 74% after first preprocessing. The RF and MNB performed well with the second preprocessing, achieving an accuracy value of 95%. However, DT had the lowest performance, with a value of 87%.

Table 2. The results of the TurkishSMS dataset

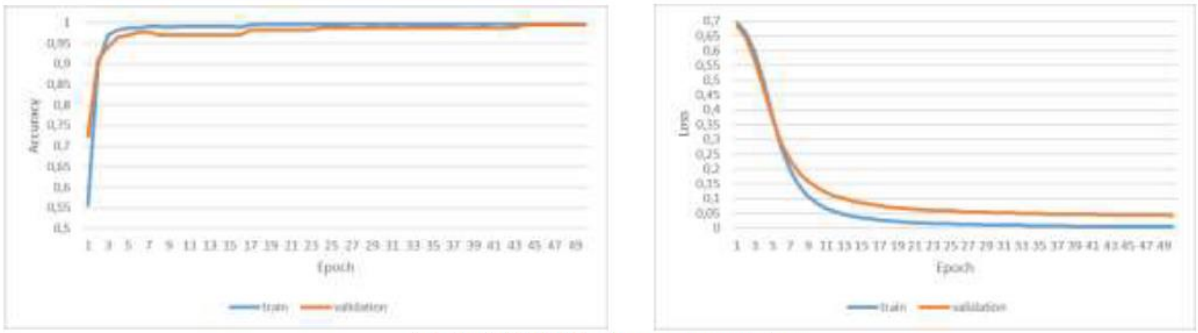
Preprocessing	First				Second			
ML Models	**Accuracy**	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
AdaBoost	0.9471	0.9487	0.9471	0.9471	0.9352	0.9353	0.9352	0.9353
DT	0.8588	0.8762	0.8588	0.8576	0.8705	0.8813	0.8705	0.8699
KNN	0.7353	0.7385	0.7353	0.7338	0.8941	0.8999	0.8941	0.8938
LR	0.9412	0.9422	0.9412	0.9412	0.9294	0.9304	0.9294	0.9294
MNB	0.9412	0.9472	0.9412	0.9409	0.9470	0.9475	0.9470	0.9470
RF	0.9412	0.9422	0.9412	0.9412	0.9470	0.9475	0.9470	0.9470
SVM	0.9471	0.9477	0.9471	0.9471	0.9411	0.9413	0.9411	0.9411
XGBoost	0.9235	0.9266	0.9235	0.9235	0.9352	0.9358	0.9352	0.9353

Table 3 presents the recall, precision, accuracy, and f-score results obtained from training the TurkishSMSCollection dataset with machine learning algorithms after the first and second preprocessings. After the first preprocessing, the SVM algorithm achieved the highest accuracy rate of 98%, while KNN had the lowest performance with an accuracy rate of 66%. After the second preprocessing, the AdaBoost algorithm achieved the highest accuracy rate of 98%, while MNB had the lowest performance with an accuracy rate of 66%.

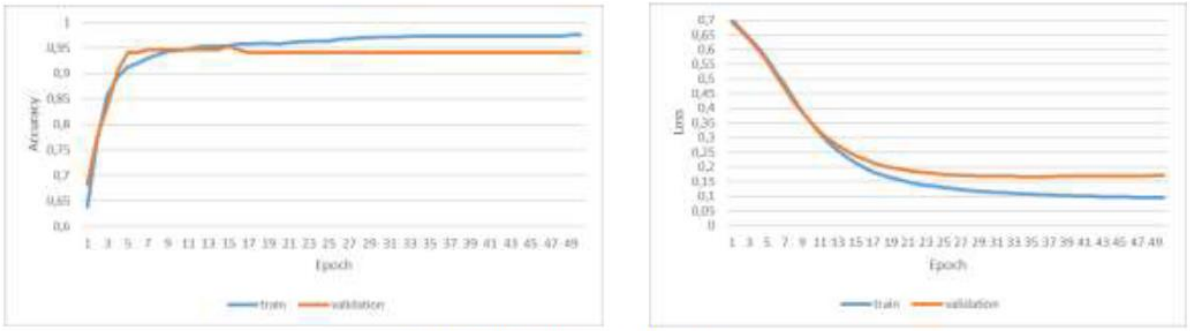
Table 3. The results of the TurkishSMSCollection dataset

Preprocessing	First				Second			
ML Models	**Accuracy**	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**
AdaBoost	0.9663	0.9664	0.9663	0.9663	0.9842	0.9842	0.9842	0.9842
DT	0.9158	0.9196	0.9158	0.9161	0.9463	0.9495	0.9463	0.9465
KNN	0.6635	0.7474	0.6635	0.6101	0.9652	0.9667	0.9652	0.9653
LR	0.9747	0.9757	0.9747	0.9748	0.9737	0.9740	0.9737	0.9737
MNB	0.9747	0.9756	0.9747	0.9746	0.6624	0.7640	0.6624	0.6039
RF	0.9284	0.9322	0.9284	0.9278	0.9821	0.9821	0.9821	0.9821
SVM	0.9831	0.9837	0.9831	0.9832	0.9821	0.9821	0.9821	0.9821
XGBoost	0.9674	0.9678	0.9674	0.9674	0.9779	0.9779	0.9779	0.9779

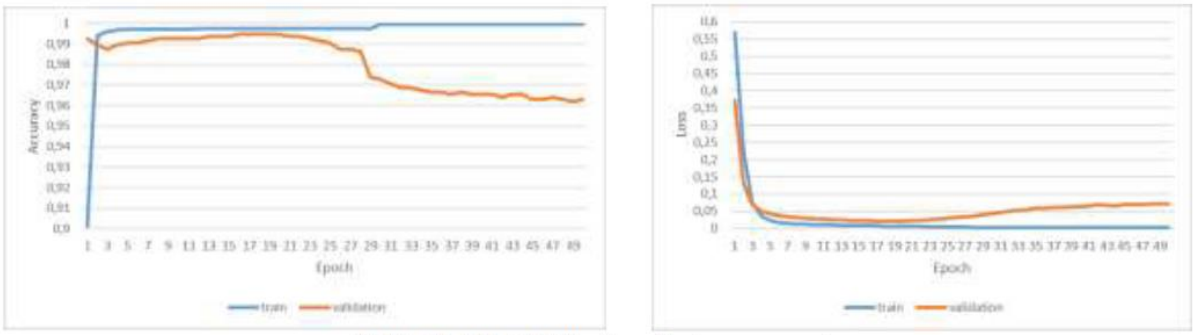
The accuracy and loss rates of the ANN with the rectified linear unit (ReLU) activation function and the adaptive moment estimation (Adam) optimizer, CNNs with the ReLU activation function and the Adam optimizer, and LSTM networks with the sigmoid activation function and the Adam optimizer are illustrated in Figures 2, 3, and 4, respectively, for both the TurkishSMS and TurkishSMSCollection datasets. All deep learning models are evaluated in a batch size of 32 and 50 epochs.



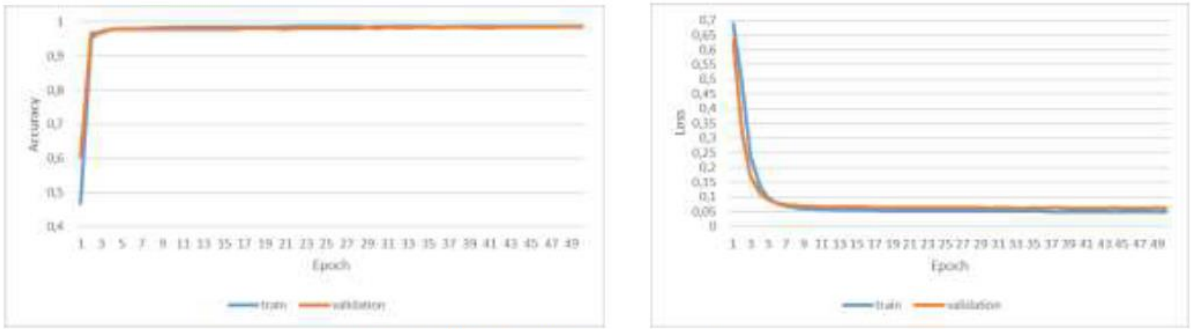
(a) TurkishSMS, first preprocessing



(b) TurkishSMS, second preprocessing

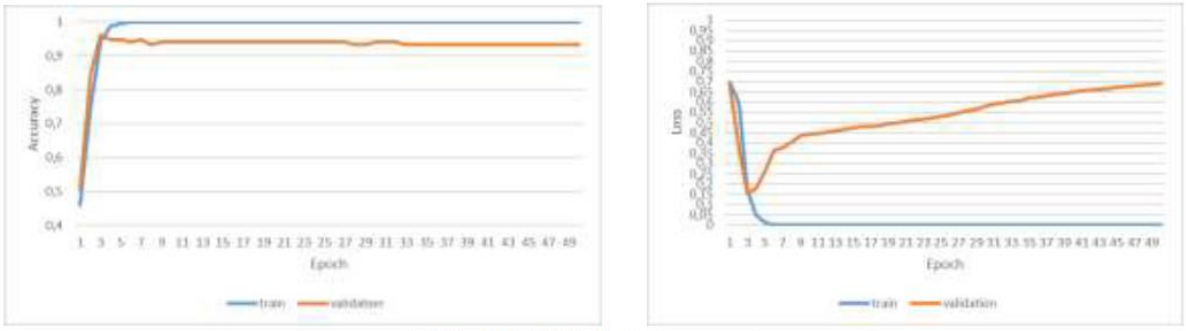


(c) TurkishSMSCollection, first preprocessing

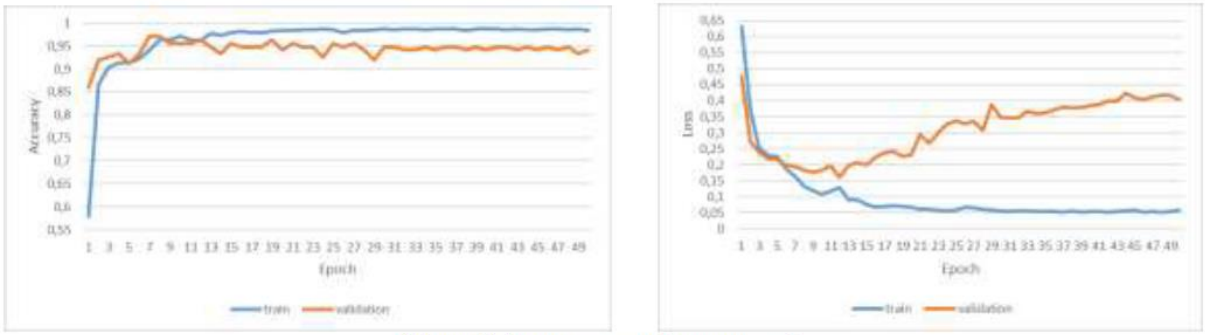


(d) TurkishSMSCollection, second preprocessing

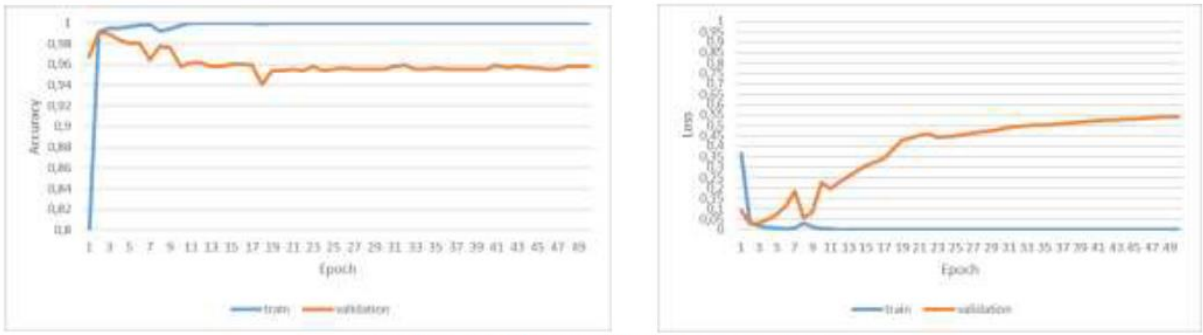
Figure 2. The accuracy and loss rate of ANN



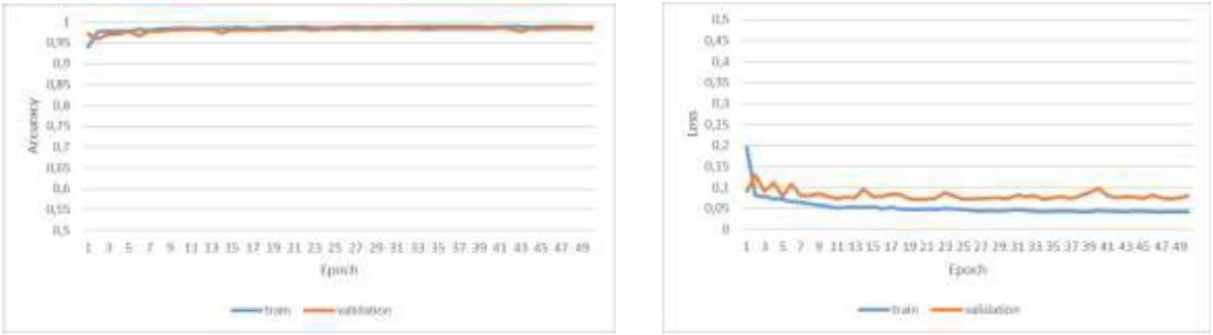
(a) TurkishSMS, first preprocessing



(b) TurkishSMS, second preprocessing

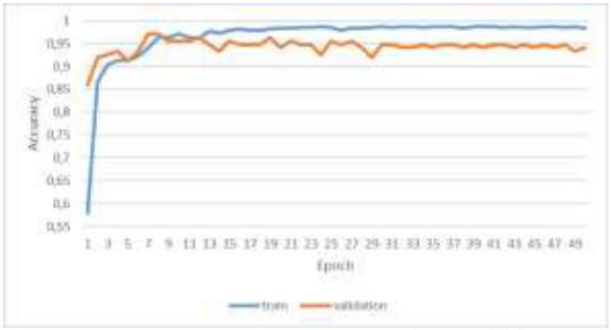


(c) TurkishSMSCollection, first preprocessing

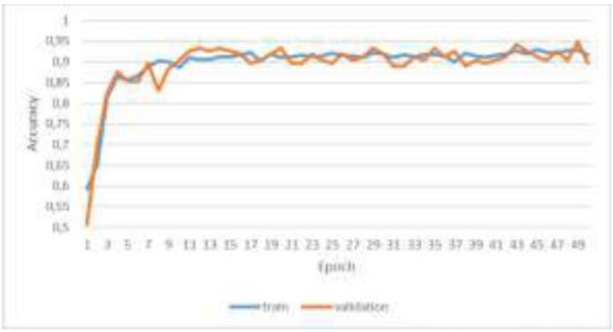


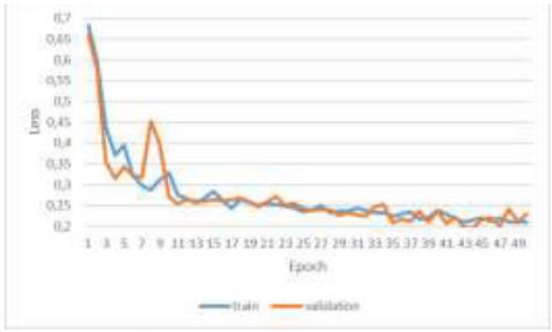
(d) TurkishSMSCollection, second preprocessing

Figure 3. The accuracy and loss rate of CNN

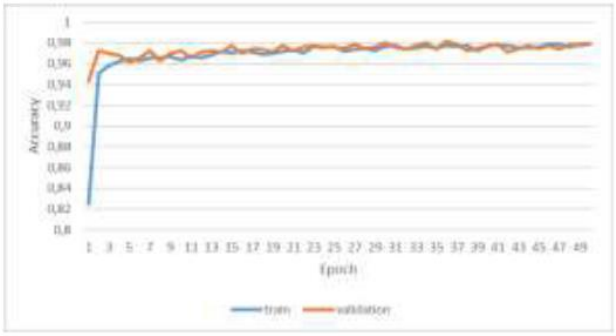


(a) TurkishSMS, first preprocessing

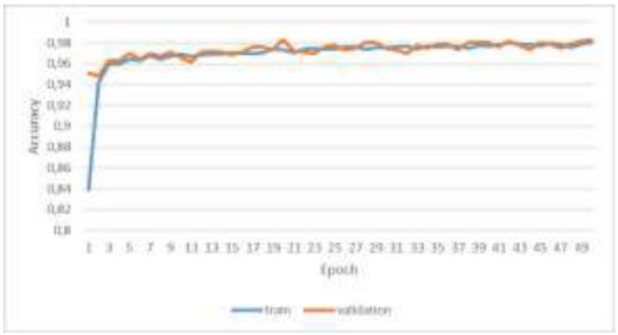


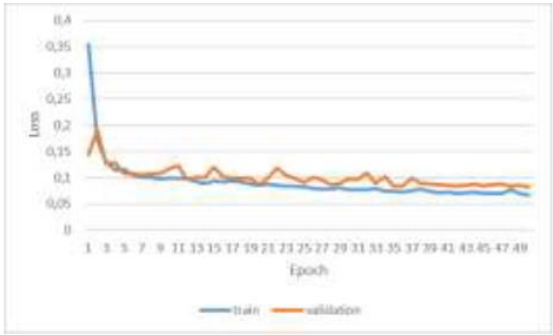


(b) TurkishSMS, second preprocessing



(c) TurkishSMSCollection, first preprocessing





(d) TurkishSMSCollection, second preprocessing

Figure 4. The accuracy and loss rate of LSTM

Cross-validation is a statistical resampling technique used to measure the efficiency of a machine learning model objectively and accurately on unseen data. It provides a more accurate idea of how the model will perform with real-world data. This study employed the non-parametric K-fold crossvalidation method. By dividing the dataset into 'k' equal parts, validation data is created and processed for each part individually. This ensures that each data point is used as validation data at least once, resulting in a more accurate evaluation of the overall performance of the model. In this study, k values of 3, 5, and 10 were tested on the TurkishSMS dataset, and the accuracy values are presented in Table 4.

Table 4. The K-fold cross validation result of algorithms on the TurkishSMS dataset

Algorithm	K = 3	K = 5	K = 10
AdaBoost	0.9561	0.9589	0.9545
KNN	0.5888	0.6426	0.6981
LR	0.9656	0.9716	0.9722
DT	0.9156	0.9147	0.9172
MNB	0.9548	0.9545	0.9676
RF	0.9672	0.9603	0.9726
SVM	0.9661	0.9710	0.9770
XGBoost	0.9406	0.9445	0.9483

As the best results were consistently achieved with a value of 10 for k, this value was selected for all algorithms. Table 5 displays the accuracy values for all algorithms resulting from two different preprocessing methods applied to the TurkishSMS, TurkishSMSCollection, and the BigTurkishSMS datasets, when k was set to 10. The TurkishSMS dataset yielded the highest value of 99.41% with ANN after the first preprocessing. The second preprocessing resulted in the highest value of 96.60% with SVM. Similarly, the SVM algorithm produced the highest score of 99.24% for the TurkishSMSCollection dataset after the first preprocessing, while the second preprocessing resulted in the highest value of 98.85% with RF.

Table 5. The accuracy results of the algorithms with 10-fold cross-validation

Dataset	TurkishSMS		TurkishSMSCollection		BigTurkishSMS	
Preprocessing	**First**	**Second**	**First**	**Second**	**First**	**Second**
AdaBoost	95.45%	94.64%	97.87%	98.48%	97.85%	97.88%
DT	91.72%	91.27%	95.73%	97.79%	94.77%	96.20%
KNN	69.81%	94.36%	64.59%	97.90%	69.42%	97.67%
LR	97.22%	96.06%	98.92%	98.78%	98.88%	98.07%
MNB	96.76%	94.72%	98.74%	65.01%	97.86%	65.06%
RF	97.26%	95.37%	89.77%	98.85%	89.72%	98.07%
SVM	97.70%	96.60%	99.24%	98.63%	99.03%	97.44%
XGBoost	94.83%	95.21%	98.02%	98.68%	97.60%	98.01%
ANN	99.41%	94.12%	96.32%	98.53%	95.63%	97.41%
CNN	95.88%	94.12%	95.79%	98.32%	95.72%	97.15%
LSTM	90.00%	92.94%	97.69%	97.79%	59.87%	96.34%

The accuracy values for all algorithms resulting from two different preprocessing methods applied to the BigTurkishSMS datasets, as seen in Table 5, with 10-fold cross-validation. The BigTurkishSMS dataset resulted in the highest accuracy of 99.03% using SVM after the first preprocessing and the highest accuracy of 98.07% using LR and RF after the second preprocessing.

The study analyzed the SMS datasets and found that the second preprocessing method resulted in eight learning models with better results for the TurkishSMSCollection and BigTurkishSMS datasets. Similarly, for the TurkishSMS dataset, the first preprocessing method resulted in eight learning models with better results. The study also observed that the second preprocessing method had better accuracy in detecting spam messages as the dataset size increased. In fact, the BigTurkishSMS dataset showed improved performance for RF, DT, AdaBoost, ANN, and CNN learning models, which did not perform well on the TurkishSMS dataset with the second preprocessing. The accuracy of spam message detection was improved by 1.0% to 3.5% with the second preprocessing.

The evaluation of the learning algorithms resulted in the following outcomes:

• In the TurkishSMS dataset, the second preprocessing outperformed the KNN, XGBoost, and LSTM models by 24.5%, 0.38%, and 2.94%, respectively. Meanwhile, the first preprocessing outperformed RF, LR, MNB, SVM, DT, AdaBoost, ANN, and CNN by 1.89%, 1.16%, 2.04%, 1.10%, 0.45%, 0.81%, 5.29%, and 1.76%, respectively.

• The second preprocessing method yielded better results in the TurkishSMSCollection dataset, with increases of 33.31%, 9.08%, 2.06%, 0.61%, 0.66%, 2.21%, 2.53%, and 0.10% in KNN, RF, DT, AdaBoost, XGBoost, ANN, CNN, and LSTM models, respectively. Conversely, the first preprocessing method resulted in better performance in LR, MNB, and SVM models, with increases of 0.14%, 33.73%, and 0.61%, respectively.

• The BigTurkishSMS dataset showed that the second preprocessing method resulted in better performance for KNN, RF, DT, AdaBoost, XGBoost, ANN, CNN, and LSTM models with increases of 28.25%, 8.35%, 1.43%, 0.03%, 0.41%, 1.78%, 1.43%, and 36.47%, respectively. On the other hand, the first preprocessing method resulted in better performance for LR, MNB, and SVM models with increases of 0.81%, 32.80%, and 1.59%, respectively.

• In the second preprocessing, KNN and LSTM outperformed their performance in the first preprocessing. Conversely, MNB performed better in the first preprocess than in the second preprocessing.

4. DISCUSSION AND CONCLUSIONS

Short message service (SMS) is a widely used form of communication with mobile devices. However, it is also a common target for social engineering scams, particularly in promotional messages from banks, betting companies, vacation providers, and similar businesses. Phishing attacks, which aim to obtain passwords, credentials, or similar data by sending fake messages to the target person, such as gifts, discounts, or similar tempting messages, are increasing day by day. To protect against such attacks, it is important to ignore messages containing unusual and untrusted URLs, requests for verification of critical personal information such as card details or passwords, errors in the language used in the message, and misspellings. Education and awareness, as well as the development of spam filtering structures, are crucial in preventing phishing attacks.

Currently, numerous studies have been conducted on spam mail and comment filtering using various learning models. Although there have been many studies on spam filtering, particularly with English real datasets, there have been fewer studies conducted with Turkish real datasets. This study utilized two significant datasets, TurkishSMS and TurkishSMSCollection, which exclusively contain Turkish spam messages. The study aimed to contribute to the field of Turkish spam filtering. By merging both datasets, a new dataset is created, named BigTurkishSMS. The datasets undergo preprocessing based on specific criteria before being used to detect spam messages. Two different preprocessing steps were used in this study. The resulting datasets were trained and tested using eight machine learning and three deep learning models. The f-score, recall, precision, and accuracy values were calculated to compare the results obtained from each model. The study aimed to provide preliminary information for future new detection models on Turkish spam. In this study, we compared the performance of eleven algorithms on three datasets using two preprocessing methods. We tested the accuracy rates of these models using 10-fold cross-validation. The highest accuracy rate of 99.41% was achieved with ANN when the first preprocessing method was used, while the highest accuracy rate of 98.85% was achieved with RF when the second preprocessing method was used. When the dataset resulting from the second preprocessing was trained with deep learning models, it increased for all datasets. Additionally, the second preprocessing resulted in an average increase in the accuracy rate of between 1.0% and 3.5% for all datasets.

According to a report by The International Business Machines Corporation (IBM), our country ranks 6th in terms of exposure to cyber-attacks. The report indicates that, following the pandemic, there was a 217 percent increase in phishing attacks and a 220 percent increase in spam attacks. This

study aimed to create more effective Turkish spam filtering models to combat phishing attacks using simple SMS. The study highlighted the importance of message filtering with learning algorithms in combating SMS phishing attacks and aimed to raise awareness on this issue. To strengthen the models used in this study, it is suggested that more recent examples of spam messages be included. Future studies could focus on developing deep learning models using new Turkish datasets.

5. REFERENCES

Al Maruf, A., Al Numan, A., Haque, M. M., Jidney, T. T., & Aung, Z. (2023, April). Ensembleapproach to classify spam SMS from Bengali text. In International Conference on Advances in Computing and Data Sciences (pp. 440-453). Cham: Springer Nature Switzerland.

Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011, September). 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).

Arulprakash, M. (2021). Eshort message service spam detection and filtering using machine learning approach. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), 721- 727.

Chen, Y. H., Huang, L., Wang, C. D., Fu, M., Huang, S. Q., Huang, J., Tan & Yan, C. (2022). Adversarial Spam Detector with Character Similarity Network. IEEE Transactions on Industrial Informatics, 19(3), 2541-2551. doi: 10.1109/TII.2022.3177726

Dierks, Z., (2023). Forecast of the smartphone user penetration rate in Turkey 2018-2024. Tech. rep., Statista.

Ergin, S., & Isik, S. (2014a, June). The assessment of feature selection methods on agglutinative language for spam email detection: A special case for Turkish. In 2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings (pp. 122-125). IEEE.

Ergin, S., & Isik, S. (2014b, June). The investigation on the effect of feature vector dimension for spam email detection with a new framework. In 2014 9th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-4). IEEE.

Eryılmaz, E. E., Şahin, D. Ö., & Kılıç, E. (2020, June). Filtering Turkish spam using LSTM from deep learning techniques. In 2020 8th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6). IEEE.

Gupta, M., Bakliwal, A., Agarwal, S., & Mehndiratta, P. (2018, August). A comparative study of spam SMS detection using machine learning classifiers. In 2018 eleventh international conference on contemporary computing (IC3) (pp. 1-7). IEEE.

Jain, G., Sharma, M., & Agarwal, B. (2019). Optimizing semantic LSTM for spam detection. International Journal of Information Technology, 11, 239-250. doi: 10.1007/s41870-018-0157-5

Karamollaoglu, H., Dogru, İ. A., & Dorterler, M. (2018, October). Detection of Spam E-mails with Machine Learning Methods. In 2018 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-5). IEEE.

Karasoy, O., & Ballı, S. (2022). Spam SMS detection for Turkish language with deep text analysis and deep learning methods. Arabian Journal for Science and Engineering, 47(8), 9361-9377. doi: 10.1007/s13369-021-06187-1

Kaya, Y., & Ertuğrul, Ö. F. (2016). A novel feature extraction approach in SMS spam filtering for mobile communication: one-dimensional ternary patterns. Security and communication networks, 9(17), 4680-4690. doi: 10.1002/sec.1660

Kemp, S., (2024). Digital 2024 global overview report. Tech. rep., Meltwater and We Are Social.

Masum, E., & Samet, R. (2018). Mobil BOTNET ile DDOS Saldırısı. Bilişim Teknolojileri Dergisi, 11(2), 111-121. doi: 10.17671/gazibtd.306612

Mathew, K., & Issac, B. (2011, December). Intelligent spam classification for mobile text message. In Proceedings of 2011 International Conference on Computer Science and Network Technology (Vol. 1, pp. 101-105). IEEE.

Matthew Shanahan, K.B., (2023). The state of mobile internet connectivity report. Tech. rep., GSMA Intelligence.

Örnek, Ö. (2019). Orange 3 ile Türkçe ve İngilizce SMS Mesajlarında Spam Tespiti. Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 1(1), 1-4.

Özdemi r, C., Ataş, M., & Özer, A. B. (2013, April). Classification of Turkish spam e-mails with artificial immune system. In 2013 21st Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

Roy, P. K., Singh, J. P., & Banerjee, S. (2020). Deep learning to filter SMS Spam. Future Generation Computer Systems, 102, 524-533. doi: 10.1016/j.future.2019.09.001

Sajedi, H., Parast, G. Z., & Akbari, F. (2016). SMS spam filtering using machine learning techniques: A survey. Machine Learning Research, 1(1), 1-14. doi: 10.11648/j.mlr.20160101.11

Salman, M., Ikram, M., & Kaafar, M. A. (2024). Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models. IEEE Access, 12, 24306–24324. doi: 10.1109/ACCESS.2024.3364671

Suleiman, D., Al-Naymat, G., & Itriq, M. (2020). Deep SMS Spam Detection using H2O Platform. International Journal of Advanced Trends in Computer Science and Engineering, 9(5), 9179–9188. doi: 10.30534/ijatcse/2020/326952020

Theodorus, A., Prasetyo, T. K., Hartono, R., & Suhartono, D. (2021, April). Short message service (SMS) spam filtering using machine learning in Bahasa Indonesia. In 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT) (pp. 199-203). IEEE.

United Nations Department of Economic and Social Affairs, Population Division (2022). World Population Prospects 2022: Summary of Results. UN DESA/POP/2022/TR/NO. 3.

Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2013). The impact of feature extraction and selection on SMS spam filtering. Elektronika ir Elektrotechnika, 19(5), 67-72. doi: 10.5755/j01.eee.19.5.1829